Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/survophthal

Review article

Automatic detection of glaucoma via fundus imaging and artificial intelligence: A review

Lauren J. Coan, MSc [a,*], *Bryan M. Williams, PhD* [b], *Venkatesh Krishna Adithya, BTech* [c], *Swati Upadhyaya, DNB* [c], *Ala Alkafri, PhD* [d], *Silvester Czanner, PhD* [a,e], *Rengaraj Venkatesh, DNB* [f], *Colin E. Willoughby, Professor* [g], *Srinivasan Kavitha, MS* [c], *Gabriela Czanner, PhD* [a,e]

[a] *School of Computer Science and Mathematics, Liverpool John Moores University, UK*
[b] *School of Computing and Communications, Lancaster University, UK*
[c] *Department of Glaucoma, Aravind Eye Hospital, Pondicherry, India*
[d] *School of Computing, Engineering & Digital Technologies, Teesside University, UK*
[e] *Faculty of Informatics and Information Technologies, Slovak University of Technology, Slovakia*
[f] *Department of Glaucoma and Chief Medical Officer, Aravind Eye Hospital, Pondicherry, India*
[g] *Biomedical Sciences Research Institute, Ulster University, UK*

ARTICLE INFO

Article history:
Received 17 January 2022
Revised 4 August 2022
Accepted 8 August 2022
Available online 17 August 2022

Keywords:
Glaucoma
Artificial intelligence
Automatic detection
Prediction
Fundus images/imaging
Classification/discrimination
Segment/segmented/segmentation

ABSTRACT

Glaucoma is a leading cause of irreversible vision impairment globally, and cases are continuously rising worldwide. Early detection is crucial, allowing timely intervention that can prevent further visual field loss. To detect glaucoma an examination of the optic nerve head via fundus imaging can be performed, at the center of which is the assessment of the optic cup and disc boundaries. Fundus imaging is noninvasive and low-cost; however, image examination relies on subjective, time-consuming, and costly expert assessments.

A timely question to ask is: “Can artificial intelligence mimic glaucoma assessments made by experts?” Specifically, can artificial intelligence automatically find the boundaries of the optic cup and disc (providing a so-called segmented fundus image) and then use the segmented image to identify glaucoma with high accuracy?

We conducted a comprehensive review on artificial intelligence-enabled glaucoma detection frameworks that produce and use segmented fundus images and summarized the advantages and disadvantages of such frameworks. We identified 36 relevant papers from 2011 to 2021 and 2 main approaches: 1) logical rule-based frameworks, based on a set of rules; and 2) machine learning/statistical modeling-based frameworks. We critically eval-

[*] Corresponding author: Lauren Coan, MSc, School of Computer Science and Mathematics, Liverpool John Moores University, Kingsway House, Crosby Road, Liverpool, Merseyside L3 2AJ, United Kingdom.
E-mail address: l.coan@2016.ljmu.ac.uk (L.J. Coan).

https://doi.org/10.1016/j.survophthal.2022.08.005

uated the state-of-art of the 2 approaches, identified gaps in the literature and pointed at areas for future research.

1. Introduction

Glaucoma is one of the leading causes of global vision impairment[A] and the second most common cause of blindness globally (86). By 2040, it is estimated that 112 million individuals globally will have the disease (90). With the aging global population (86), there will be a corresponding increase in glaucoma cases that will continuously challenge our resources worldwide (76). The global burden of vision impairment and/or blindness from glaucoma is significantly associated with a decrease in quality of life, physical functioning, and mental health (22). Although irreversible, early diagnosis of glaucomatous optic neuropathy allows for treatment to be implemented that may slow or prevent glaucoma progression and blindness.

Currently, in the United Kingdom (UK), glaucoma detection is opportunistic, most frequently accomplished by optometrist assessment in the community (42). Around half of the glaucoma patients in the community remain undiagnosed (16). A recent population-based study in Northern Ireland suggests that the majority of people with glaucoma are undetected, and two-thirds of glaucoma patients within the study were unaware of having the disease (63).

Although a worldwide problem, the burden of glaucoma is higher within developing countries (30), and the disease disproportionately affects African and Asian countries (78). Moreover, studies indicate that more than 11.2 million individuals in India are affected by glaucoma, constituting approximately one-fifth of the global burden of the disease (81). In the UK, hospital eye services (HES) are the busiest outpatient service in the National Health System (NHS) and are responsible for 8.3% of all outpatient activity[B]. Glaucoma accounts for 25% of HES appointments. Individuals with, or at risk of, glaucoma are detected by community optometrists and referred to HES, 15%–20% of the new referrals will have glaucoma and around 50% will be discharged at the first visit, costing the NHS upwards of £75m/year (10).

Given this worldwide problem of glaucoma detection, the urgent question is how close we are to having accurate artificial intelligence (AI)-enabled glaucoma detection (42) and whether such AI can then be explained to the clinician and patient. The answer to this question is two-fold: we need to understand the process of detecting glaucoma in clinical practice, and then we need to determine if artificial intelligence can accurately detect glaucoma while also providing key explanations, mimicking the clinician's reasoning.

1.1. The detection of glaucoma by a clinician

Glaucoma is a chronic progressive optic neuropathy in which changes in the structure of the optic nerve head (ONH) (Fig. 1A) and retinal nerve fiber layer (RNFL) are associated with visual defects. Structural changes are manifested by a slow, yet progressive, narrowing of the neuroretinal rim, indicating degeneration of retinal ganglion cells axons, and astrocytes of the optic nerve (13). To evaluate the narrowing of the neuroretinal rim (NRR) the clinician needs to identify the boundary contours of the cup and disc. Such contours then help when explaining to the patient the reasoning behind the diagnosis, and thus help the patient to participate in the discussion and treatment decision. Given the significance of patient involvement in the decisions regarding their care and the importance of AI explainability, this review focuses on AI that provides optic cups and optic disc contours.

Glaucoma detection is a challenging and lengthy process, relying on multiple examinations and clinical expertise. The National Institute for Health and Care Excellence (NICE) in the UK recommends examination of the ONH via a technique called fundus imaging (22). Imaging modalities are key for evaluating structural abnormalities in the ONH. Such structural abnormalities often precede the development of visual field loss (87).

One method of fundus imaging is color images collected by fundus cameras (Fig. 1).

Another fundus imaging technology is optical coherence tomography (OCT), which can provide 3-dimensional information to aid glaucoma diagnosis. The interpretation of color fundus image vs OCT is different, though both essentially evaluate the structure of ONH. OCT outputs provide numerical and graphical representations of the peripapillary retinal nerve fiber layer compared to age-matched normative data in an objective way. A report can be generated from this output (dependent on the OCT platform used). This report assists clinicians in the interpretation and the identification of glaucoma-related abnormalities thus OCT can require less clinical expertise than the interpretation of a color optic disc image.

Fundus cameras are advantageous owing to their relatively low cost compared to their imaging counterparts such as OCT and Heidelberg Retinal Tomography (HRT). Yet, they provide images that are of suitable quality to detect abnormalities in the ONH for evaluating ocular health (92). Owing to their cheaper cost, fundus cameras are readily available in a range of settings including rural community centers, local ophthalmologist offices, and hospitals. Although in recent years, OCT imaging has become cheaper and more widely available to optometrists in economically stronger countries. Low-cost portable fundus cameras have been developed that can be more readily utilized for wider population-based screening of glaucoma in lower resource settings or isolated communities.

Portable fundus cameras are becoming increasingly accessible and viable (11) even within economically less fortunate countries. These recently developed smaller mobile cameras enable high-quality imaging of the ONH at a considerably

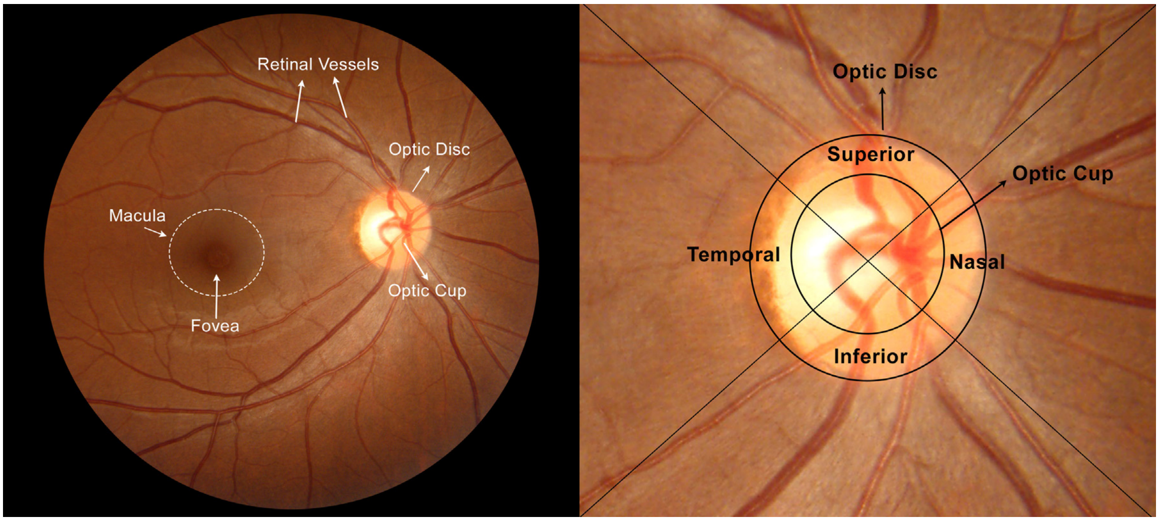

Fig. 1 – Fundus photograph examples (A- left) with labels of the optic nerve head and (B - right) with (Inferior-Superior-Nasal-Temporal) ISNT quadrants.

lower cost, providing a more cost-effective alternative to tabletop devices (26). Potentially, portable fundus cameras can be used to identify suspects in glaucoma screening programs, outside of the hospital setting (communities or optometry centers). Once an individual is suspected to have glaucoma based on the fundus imaging, they must undergo a comprehensive glaucoma evaluation including an assessment of visual acuity, IOP, gonioscopy, and visual fields. Therefore, this review focuses on AI that utilizes fundus images.

1.2. *Detecting glaucoma via artificial intelligence (AI)*

AI is a computer system that can perform tasks that normally require human intelligence such as glaucoma detection via examination of fundus images. AI methods are developed by applying technical expertise (in data science, mathematics, and computing – also known as algorithmic expertise) to interrogate the data, which leads to producing fast and intelligent computer algorithms. Often, but not always, human intelligence (such as knowledge of rim thinning in glaucoma) is also applied in synergy with algorithmic intelligence. AI is an umbrella term that encapsulates machine learning algorithms, which in turn include deep learning (DL) methods. In recent years we have seen a significant increase in the utilization and development of AI, alongside momentous developments in technology[C]. Automated algorithms are already being used in some clinics including ophthalmology (7) such as the FDA-approved AI-based device that detects diabetic retinopathy[D].

Technological advances mean that the creation of AI-enabled glaucoma detection methods via the modality of fundus images is a realistic proposition (68). Several portable fundus cameras have been developed; such devices are small, inexpensive, and are becoming straightforward enough to be operated by laypersons (48). A recent review on the use of telemedicine in glaucoma highlights that machines that are less operator-dependent should give more objective results even when they are operated by less experienced personnel at remote sites (56).

If AI-enabled glaucoma detection methods using fundus imaging could be deployed in screening mechanisms, this could aid in reducing human error (e.g., observer bias and fatigue) and be used for large-scale screening at a low cost. This could provide much-needed eye care services to remote rural areas, particularly in nations where there is a scarcity of qualified, skilled, and competent ophthalmologists (60). In the near future, automated image interpretation for screening, referral decision-making, and patient monitoring is likely to play a crucial role in frontline eye care. Even in resource-rich care settings such as the NHS in the UK, referral refinement with AI has the potential to address the staggering outpatient appointment demand while reducing false positive referral rates[E].

What remains unclear is the full state of AI-enabled glaucoma detection, namely the frameworks that utilize fundus cameras while providing the contours of the optic cup and disc. To understand the potential application of AI-enabled glaucoma detection, we must first answer many questions (i.e., how accurate are the AI methods, how suitable/appropriate are they, and how have they been trained/tested/validated). Following this, we can then identify the next steps to further develop AI-enabled glaucoma detection.

1.3. *There are two AI approaches for glaucoma detection*

AI for glaucoma detection can be split into two approaches: one-step and two-step. In a one-step approach, the AI detects glaucoma in a single step. The only way to do it is via deep learning black-box approaches, also called end-to-end approaches. The two-step approach to glaucoma detection is to proceed in 2 steps. In the first step, AI can be applied to find the optic cup and disc contours, then a second step uses the information from the first step for the derivation of the automated decision rule for glaucoma detection. One-step approaches do not find nor provide the contours of the optic cup and optic disc (i.e., they do not provide segmentation).

This review solely focuses on two-step AI approaches for two primary reasons. Firstly, two-step AI approaches may have advantages over the one-step approaches that are unknown to the AI community at large. Secondly, reviews of solely two-step approaches are absent from the literature. Previous reviews have already extensively covered one-step/end-to-end approaches see (2,17,64,91). A detailed comparison of the two approaches is in Section 3.6.

1.4. *Overview*

Our key objectives are: (1) to outline and clarify the main AI terminology used with AI-enabled glaucoma detection such that the review is accessible to ophthalmologists, and (2) to provide a detailed overview of the state-of-art AI-enabled glaucoma detection methods that use segmented fundus images - highlighting the two approaches used when using fundus imaging, and (3) to provide a discussion on the progress of AI-enabled glaucoma detection methods and highlight areas that require further work.

In the following sections, we provide a clinical and technical background and define the terminology referred to throughout this review. Section 3 then defines the methods used for the literature search and outlines the key information extracted from the reviewed papers. Section 4 explains the methods employed in this review and Section 5 covers the results of the review. Lastly, Section 6 provides a discussion, conclusions, and future work recommendations.

2. Clinical terminology and brief background

2.1. *Cup-to-disc ratio*

The cup-to-disc ratio (CDR) is a universally acknowledged parameter for describing glaucomatous neuropathy, obtained from assessment of the ONH. There are different variants of the CDR parameter however, the primary two are the vertical cup-to-disc ratio (vCDR) and the area cup-to-disc ratio ACDR.

The vCDR is defined as:

$$vCDR = \frac{Vertical\ Cup\ Diameter}{Vertical\ Disc\ Diameter}$$

The ACDR is defined as:

$$ACDR = \frac{Area\ of\ Cup}{Area\ of\ Disc}$$

Although well used in practice, the CDR parameter is limited in cases of genetically large or small optic disc, large optic cup cases, and in cases where myopic ONH changes are present (65,21); in such instances, the CDR can be misleading (41) and lead to errors in diagnosis. Other morphometric features such as the rim-to-disc ratio (RDR) and horizontal cup-to-disc ratio (hCDR) can also be considered. In contrast to the CDR, a decrease in the RDR indicates glaucomatous neuropathy. The ACDR provides a 2-D feature-based measurement allowing structural changes of the ONH to be assessed.

2.2. *Neuroretinal rim area ratio*

nThe reuroretinal im (NRR) is the area between the optic cup margin and the optic disc margin which comprises retinal nerve fiber axons. When using fundus images, the NRR is the area left behind when subtracting the optic cup from the disc. The NRR is divided into four quadrants: inferior, superior, nasal, and temporal as shown in Fig. 1B.

The NRR area (89) is calculated as:

$$NRR = \frac{Area\ in\ Inferior\ Quadrant + Area\ in\ Superior\ Quadrant}{Area\ in\ Nasal\ Quadrant + Area\ in\ Temporal\ Quadrant}$$

The four quadrants of the NRR are typically expected to satisfy the inferior-superior-nasal-temporal (ISNT) rule (I>S>N>T) (65). Whilst the cup-to-disc ratio parameter focuses on the optic cup size with respect to the optic disc, the ISNT rule focuses on the NRR width i.e., the area between the boundary of the optic cup and disc (65). The ISNT rule follows that the inferior rim is thicker than the superior rim, which is thicker than the nasal rim, which is thicker than the temporal rim in a healthy eye (24). Any violation of the ISNT rule can be seen as a sign of glaucomatous neuropathy. However, this is not always the case (i.e., a healthy NRR can violate the rule) (89). As such, the ISNT rule is not recognized as a diagnostic test, but rather a clinical tool.

2.3. *Disc damage likelihood scale*

The Disc Damage Likelihood Scale (DDLS) is a grading protocol that divides glaucomatous progression into 10 stages while accounting for optic disc size (85). The advantage of this method is in higher inter-observer repeatability (87) and higher agreement with the gold standard (21) than the vertical CDR. The DDLS method has proved to be time-consuming, requiring a detailed grading protocol with a standard set of images for comparison purposes. Also, it necessitates further training of clinicians.

3. Technical terminology and brief background

Within the AI community, many terms are used interchangeably; we define the key terminology used throughout this review.

3.1. *Fundus image segmentation*

In medical image processing, image segmentation refers to the (typically automated) partitioning of an image into multiple clinically meaningful segments (Fig. 2). Fundus image segmentation is the process of finding the visible boundaries (or "contours") of the optic cup and disc. Manual image segmentation can involve a trained expert, such as a clinician or grader, manually annotating the boundary of the optic cup and disc. Whereas automatic image segmentation is accomplished by mathematical algorithms. To date, there have been a large number of AI methods proposed for automatic image segmentation of the ONH. Popular approaches include level-set-based

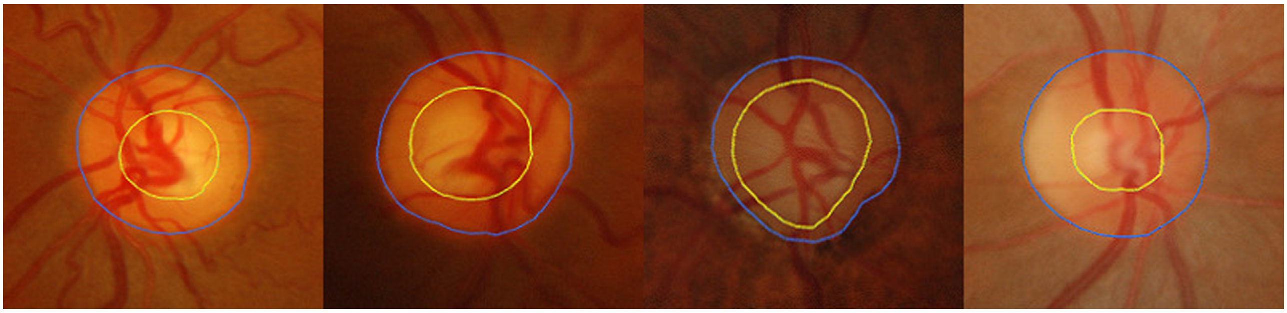

Fig. 2 – Examples of the automatic optic cup and disc segmentation in fundus images centered on the optic nerve head. The yellow contour represents the optic cup boundary, and the blue contour represents the optic disc boundary.

algorithms, threshold-based algorithms, and clustering-based algorithms (9,94). The resulting annotation of the boundaries is what we call the *segmented image*.

3.2. *Image features*

In the AI community, the term "image feature" refers to a variable or parameter derived from an image. Two types of image features can be extracted from fundus images: namely, clinically interpretable features and abstract features.

Clinically interpretable features are features with clinical meaning (e.g., vCDR and NRR area). These clinical features have been developed over many years by expert ophthalmologists and can be intuitively explained to a patient. In contrast, we can also consider mathematically derived *abstract features*. Such features may not be clinically interpretable as they are constructed via a mathematical or statistical process.

3.3. *Probability of glaucoma*

In general, AI calculates the probability of glaucoma for an unseen new fundus image as a number between 0 and 100% (e.g., 90%). This probability is interpreted as follows: given the training set that AI used and the mathematical/statistical method that the AI is built on, the AI believes that the chance of glaucoma is 90%, i.e., among the 10 images that look like the new image, 9 do have glaucoma and 1 does not. The value of the probability of glaucoma should be calculated to reflect the prevalence in the population of interest via e.g., Bayesian updating rule. If the probability provided by AI is 50%, then the AI is not certain if the new image is glaucomatous or not; however, if the probability is 99%, this does not mean that AI is certain that it is glaucoma. The probability estimates provided by AI (e.g., produced by softmax or by statistical predictive algorithms) need to be calibrated to be clinically meaningful (see e.g., 1,93), as well as uncertainty needs to be ascribed to the probability estimates produced by AI models. For example, if the new image is not represented well in the training dataset, then AI is not sufficiently trained to judge the new image, and therefore it should be able to express its uncertainty[F]. The calculation of uncertainty of AI is a complex problem and is a current area of intensive research.

3.4. *Image classification*

We use the term "image classification" to refer to the automated process of determining the category to which a given fundus image belongs e.g., healthy, or glaucomatous group (binary classification); or healthy, suspected glaucoma or glaucoma group (multi-class classification). This process is also referred to as image discrimination (23) or disease prediction. To achieve the classification, AI can apply a threshold to the estimated probability of glaucoma, e.g., if the image's estimated probability is higher than the threshold, the image is classified as glaucoma. If AI is uncertain in the calculated probability, then such uncertainty will propagate into the uncertainty of the classification.

3.5. *Classifier*

We use the term "classifier" to refer to a mathematical or statistical or machine learning method used within the AI framework to estimate which disease category the patient belongs to (glaucomatous, suspected glaucomatous, or healthy). Popular classifiers are support vector machines and logistic regression.

3.6. *AI framework*

We use the term "AI framework" to encapsulate the whole process of automatically classifying a given fundus image into a group (glaucomatous, suspected glaucomatous, or healthy). This process can comprise many steps including (but not limited to) image segmentation, feature extraction, and using the image features (via various methods) for discrimination of glaucomatous neuropathy. The framework's final step is to provide the classification output for a given image.

One-step AI framework. Some AI frameworks do not require and do not produce segmented images. They learn a link between the fundus images and the disease status and then directly provide their estimate of the disease group. To build such AI, a so-called *end-to-end image classification* method is needed. Such computation can be enabled via DL algorithms (58) (e.g., convolution neural networks). This is possible due to their complex interior working architectures with complex transformations across multiple layers.

Two-step AI framework. Other AI frameworks produce a segmented image as the first step. In this step, the segmented image can provide clinically interpretable features (e.g., CDR ratio and NRR area), or abstract features (e.g., texture and color features). The second step then uses such features and provides an estimate of the disease group. In general, these two-step frameworks have increased interpretability as they have the potential to provide the clinician and patient with the segmented image, which allows demonstration of the part of the image leading to the AI's output for a patient (50) and facilitates further investigation. The concept of a two-step AI framework is not new. One recent example is the work of De Fauw in 2018 (27) for diagnosis and referral of retinal disease, however, their work does not include glaucoma.

One of the criteria by which one-step and two-step AI methods are compared is interpretability. This is one of the key elements of building trust, especially in high stake scenarios such as disease detection. Interpretability means that AI can explain its conclusion about a patient, i.e., what part of the image was most crucial in the conclusion and why AI has provided the respective output[CF]. This is related to GDPR Article 15, which stipulates that individuals have the right to access their data[G]. This includes an obligation for the controller to provide meaningful information about the logic involved and the significance and envisaged consequences of processing the individual's data via AI[H]. The principles outlined by the High-level expert group on AI appointed by the European Commission (HLEG)[I] state that it should be possible to demand a suitable explanation of the AI system's decision-making process. Not only does this impact the patient but it also puts responsibility onto the controller (i.e., the clinicians implementing the AI) to quantify and fully understand the AI to provide such information to the requesting individuals. More discussion on desired AI properties can be found in (80, [F,I,J]).

Advantages of two-step AI frameworks:

1) At the interface of the two steps, the boundaries of the optic disc and cup are provided. This enables clinicians access to intermediate representation that illustrates which part of the rim is narrowing and thus suggesting the presence of glaucoma (interpretability). This can be integrated into clinical workflows and AI quality monitoring. This can be interrogated by human experts if they want to see why a recommendation has been made. This means that clinicians can remain in the process of making a diagnosis. Such knowledge is advantageous for patients too, as it shows the areas of narrowing of the optic rim and then this offers the possibility for a patient to appeal the output of AI, as well as a possibility to participate in shaping AI design and operation. All mentioned points are crucial for building trust. Additionally, there is a utility of fundus images beyond the optic disc for glaucoma diagnosis (45).
2) Two-step frameworks may require smaller datasets for training than one-step frameworks that utilize DL. This statement is supported by the following points. Firstly, the fact that two-step methods may need less data can be explained by looking at the architecture of the AI. One-step approaches that use a DL architecture learn via complex multi-level representation transformations across many layers, with large numbers of parameters to estimate. These transformations are non-linear and are not designed manually but learned via the training data. That is, the network learns by examples, finding its own way of discerning between ground truth labels (i.e., glaucoma vs healthy). As a result, they require vast amounts of data to learn such patterns. Although in recent years we have seen a 'rise of data' there is still not an abundance of high-quality accessible data within the field of glaucoma. This is even more problematic when requiring data with high-quality annotations (ground truth) and a good sample of examples (patients, cohorts, imaging devices, etc.). This can be highlighted by the example of two proposed works. The two-step framework for glaucoma detection by MacCormick and coworkers (62) achieved an accuracy of AUROC 99.6% and 91.0%, in internal and external validation respectively, while using approximately 300 images for training. Whilst a one-step DL framework proposed by Li and coworkers (57) achieved comparable accuracy but required 30,000 images for training.

Secondly, the one-step DL approach must address the issue of dealing with lots of redundancy in the data, and a small set of labels assigned to the whole image mean that little ground-truth information is made available. The use of the whole fundus image in DL methods means that the methods have a large amount of data to handle, much of which may be redundant – with the most important information appearing to lie in the boundaries of the ONH. Thirdly, in areas outside of ophthalmology, it has also been observed that neural networks can be made more data-efficient if they utilize contours (38).

Disadvantages of two-step AI frameworks:

1) They are prone to compound errors. This is due to the sequential nature of two-step frameworks – it inherently gives rise to compounding errors. An error in the first stage of segmentation will then transpire throughout the framework and could lead to errors in the second stage and incorrect predictions. In model training, it is possible to use this as a tool for improvement. The AI developer can evaluate the AI performance in isolation (i.e., segmentation and classification performance). They can further explore any misclassifications that occur and work back to deconstruct why these are happening (i.e., segmentation error) and implement methods to improve upon this. One-step frameworks do not directly have the capacity to be interrogated in such ways but are not at the same risk of compound errors.
2) They require more domain expertise and more time for technical work for AI model development. Firstly, even though they need less training data, such training data need more clinical annotations (i.e., annotations of the boundaries of cup and disc). Secondly, the clinical knowledge needs to be elicited and then used to craft the AI model (e.g., knowledge about rim thinning in glaucoma). Thirdly, the technical team needs to find ways to incorporate the knowledge into the AI model, thus more time is needed for AI development. This all

enables increased interpretability, as well as lowers the need for vast amounts of training data (see Advantages 1 and 2).

Further comments on two-step vs one-step AI frameworks for glaucoma detection:

1) We previously highlighted (Advantage 1) that two-step AI frameworks can be constructed to facilitate explanation of the final decision, i.e., they are interpretable by design. Such frameworks are able to explain why they arrived at a conclusion that an eye has glaucoma. In contrast, the one-step frameworks relying on black-box approaches, such as DL, do not provide an explanation without post hoc descriptive methodology; however, recently the AI community is working on bringing interpretability to DL. This remains an ongoing and active research area. The interpretability of DL is being researched in two ways. Firstly, there is a research effort to make DL interpretable by design. Examples are in detecting bird species and car models (20), or text classification (19). Such methods have not been implemented for glaucoma detection. Secondly, there is an intention to develop a 'post-hoc interpretability' for DL as an additional analysis. Here, one interprets a trained DL method by fitting explanations as to how it performed the classification. This can be then visualized (i.e., saliency maps). One can find regions of the image that led to the classification output (i.e., opening the black box). Yet, whilst such post-hoc methods can aid an expert user to understand what data is most relevant to how the AI works, it provides limited insight into how that information is used. It should be noted, this underlying requirement of interpretable and explainable AI does not have to come at the cost of accurate AI (79).
2) Two-step AI frameworks may be easier to generalize and are less prone to overfitting issues than the one-step methods If AI has been 'over-fit' to specific training data, then the AI cannot be used reliably to make conclusions on future data, i.e., it lacks generalizability. The problem of overfitting can be mitigated to a degree for one-step frameworks that utilize DL with techniques such as dropout, early stopping, and regularization yet each technique has its drawbacks and overfitting remains an issue in many approaches.
3) Two-step AI frameworks may be less computationally intensive than one-step AI frameworks i.e., they need lower computational power. However, the computational intensity is (to some extent) mitigated for DL via state-of-the-art computational algorithms and hardware.

3.7. *Evaluating the performance of AI*

Careful evaluation of AI is required to understand the AI's performance capabilities; that is, how well the AI agrees with the gold standard. By the "gold standard" (also referred to as ground truth), we refer to the decision of a clinical expert on whether the eye has glaucoma or not. There is no single measure that alone would be enough to evaluate the performance of AI. Hence, a combination of measures is required to give a complete overview of the AI framework's capabilities. In what follows we briefly mention the most important measures for evaluating the performance of AI.

Confusion matrix. The confusion matrix (Table 1) is used to give an overall representation of the performance of the AI's framework. Using this confusion matrix, key performance metrics are derived.

The true positives (T_P) are the glaucomatous observations that have been correctly classified, whereas the true negatives (T_N) are the non-glaucomatous observations that are correctly classified as non-glaucomatous. The false positives (F_P) are the non-glaucomatous observations that are incorrectly classified as glaucomatous, and the false negatives (F_N) are the glaucomatous observations that are incorrectly classified as non-glaucomatous.

The accuracy metric is the proportion of correctly classified images. Sensitivity (aka true positive rate) is the proportion of actual positive cases (i.e., glaucomatous) that are classified as positive. Specificity (aka recall) is the proportion of actual negative cases (i.e., healthy) which are classified as negative.

The positive predictive value (PPV) is the probability that an individual with a positive reference test truly has the disease whilst the negative predictive value (NPV) is the probability that an individual with a negative reference test truly does not have the disease.

$$Sensitivity = \frac{T_P}{T_P + F_N}$$

$$Specificity = \frac{T_N}{T_N + F_P}$$

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}$$

$$Positive\ Predictive\ Value\ (PPV) = \frac{T_P}{T_P + F_P}$$

$$Negative\ Predictive\ Value\ (NPV) = \frac{T_N}{T_N + F_N}$$

False positives are mistakes that potentially could lead to unnecessary further testing/referrals. Arguably false negatives are more serious in glaucoma as the disease is not identified and treated at the earliest stage. The detection of glaucoma would then occur at later stages, resulting in advanced and irreversible ONH damage and possible visual field loss, impacting the patient significantly. To this end, an effective framework (with high sensitivity) for the detection of potential glaucomatous subjects at the earliest stage is paramount.

Area under receiver operating characteristic curve (AUROC). A Receiver Operating Characteristic (ROC) curve plots the true positive rate (*sensitivity*) vs the false positive rate ($1 - specificity$) at all classification thresholds. The AUROC is defined as the area under the ROC curve. If we are presented with a pair of eyes, one with glaucoma and one without glaucoma, then the AUROC metric is interpreted as the probability of correctly distinguishing the glaucomatous eye from the non-glaucomatous eye. An AUROC of 0.5 is the equivalent to the flip of a coin.

Table 1 – Confusion matrix.

		Predicted Class	
		Negative (0)	Positive (1)
Actual Class	Negative (0)	True Negative (T_N)	False Positive (F_P)
	Positive (1)	False Negative (F_N)	True Positive (T_P)

Internal and external evaluation of AI. AI methods are tested to compute the aforementioned performance metrics (i.e., accuracy, sensitivity, AUROC, etc.) AI must be evaluated on data that have not been used within its training component. There are two methods for evaluating AI: internally and externally. In internal evaluation, the dataset can be split into two partitions, one is used for training and one for testing (e.g., 70:30 split). Hence, an image can either be in the training or testing set, but not in both.

Another approach to internal evaluation is k-fold cross-validation. When using k-fold cross-validation, the dataset is randomly split into k equally sized partitions; $(k-1)$ partitions are used for training the classifier and the final partition is used for testing. This is repeated k times with the performance metrics being retained each time. The final metric presented is the average of the k splits. Generally, the value of k is set to five or 10 for optimal bias-variance trade-off (44). Such evaluation approaches are called internal as all images come from the same source (i.e., the same cohort), and hence it may not be sufficient for evaluating the generalizability of the AI.

Conversely, external evaluation consists of testing the framework on data from a different source (then the data used for training). This could be a dataset acquired from a different cohort or device. Whilst internal testing gives insight into the performance capabilities of the framework, external testing is required as it provides an understanding of the generalizability of the framework with unseen data from different sources.

3.8. *Reporting guidelines for AI in healthcare*

With the ongoing developments of AI for health applications, there has been an increase in published guidelines for the reporting of the methods. The key information that should be reported includes the imaging device, contextual study setting, detailed cohort information and data processing methods(33). With the use of AI, further detail is required to be reported comprising the technical aspects of the methods presented. Recently, new standards specific to reporting studies of machine learning/AI interventions have been in development. This includes TRIPOD-ML, SPIRIT-AI and CONSORT AI (33) under the EQUATOR initiative ^{K}.

4. Methods

We performed a comprehensive literature search, details of which can be found in the Method of Literature Search section. A table was used to extract all relevant information from the selected papers. For this review, we extracted information regarding the authors, year of publication, approach to classification, data used (sample size, availability of the data publicly, number of data annotators, imaging device details), techniques used for segmentation, validation techniques applied, performance metrics of the methods (accuracy, sensitivity, specificity and AUROC). The key terms were agreed upon by a collection of professionals with a range of experiences. This included mathematicians/statisticians and experienced clinicians. Two people reviewed titles and abstracts (LC and GC), and any disagreements were reconciled via consulting with a third person (BW). While this review is primarily focused on assessing the classification of glaucoma following segmentation, we do provide details about methods for segmentation as this is a key step in the pipeline and can heavily influence classification results.

5. Results

5.1. *Papers included*

We identified a total of 1080 papers (Fig. 3) to meet the keyword search (Section 7). After the removal of 252 duplicates, papers were screened based on titles and abstracts. A total of 623 papers were removed following title and abstract screening due to unsuitability for this review. The remaining 205 papers were screened based on text. Of these, 169 papers were removed due to unsuitability for this review, leaving 36. There were 3 main reasons papers were labeled as unsuitable in this review (from most prevalent): (1) they proposed a one-step AI framework that did not require any segmentation of the fundus images, (2) they focused purely on segmentation and provided no framework for classification of glaucomatous optic neuropathy, (3) they did not present a 2-step approach with fundus images. A total of 63 papers were identified in 2021. From the 21 papers collected for full-text reading, 5 papers were excluded-for not using segmentation, 4 were excluded as they proposed no classification (only segmentation), 1 was excluded for using solely OCT, and 1 was excluded for unclear reporting. The final number of papers that met eligibility criteria (Section 7) for this review was 36.

5.2. *Characteristics of identified papers*

We have highlighted two distinct approaches to the classification of glaucoma from segmented images. We termed the first approach the logical rule-based framework due to the use of straightforward threshold rules (IF-ELSE statements) based on clinically interpretable imaging features. The second is machine learning/statistical modeling frameworks which exploit the imaging features in a range of classification mod-

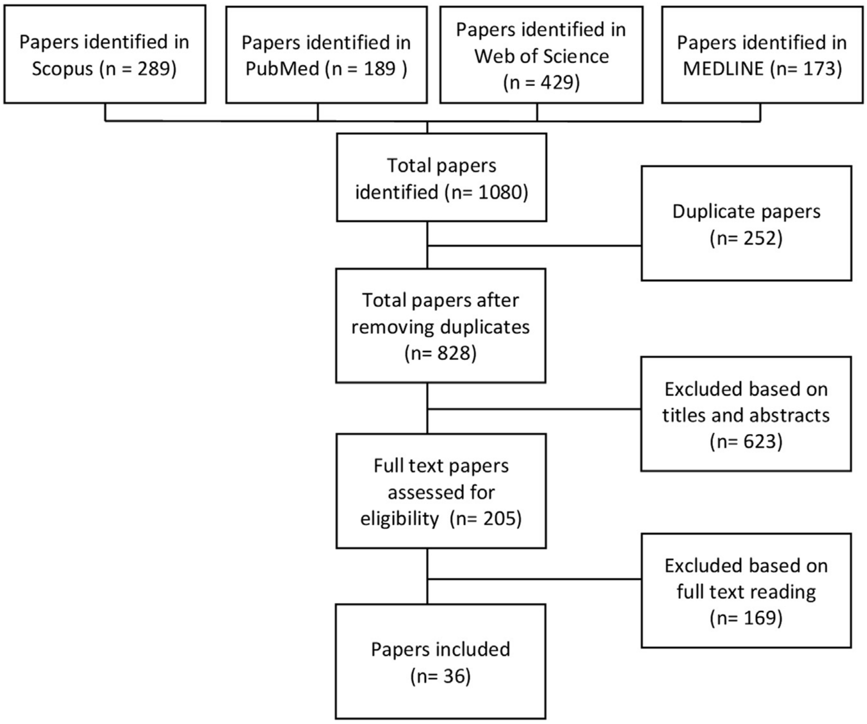
Papers identified in Scopus (n = 289)
Papers identified in PubMed (n = 189)
Papers identified in Web of Science (n = 429)
Papers identified in MEDLINE (n= 173)
Total papers identified (n= 1080)
Duplicate papers (n= 252)
Total papers after removing duplicates (n= 828)
Excluded based on titles and abstracts (n= 623)
Full text papers assessed for eligibility (n= 205)
Excluded based on full text reading (n= 169)
Papers included (n= 36)

Fig. 3 – Flow diagram of papers included within the review.

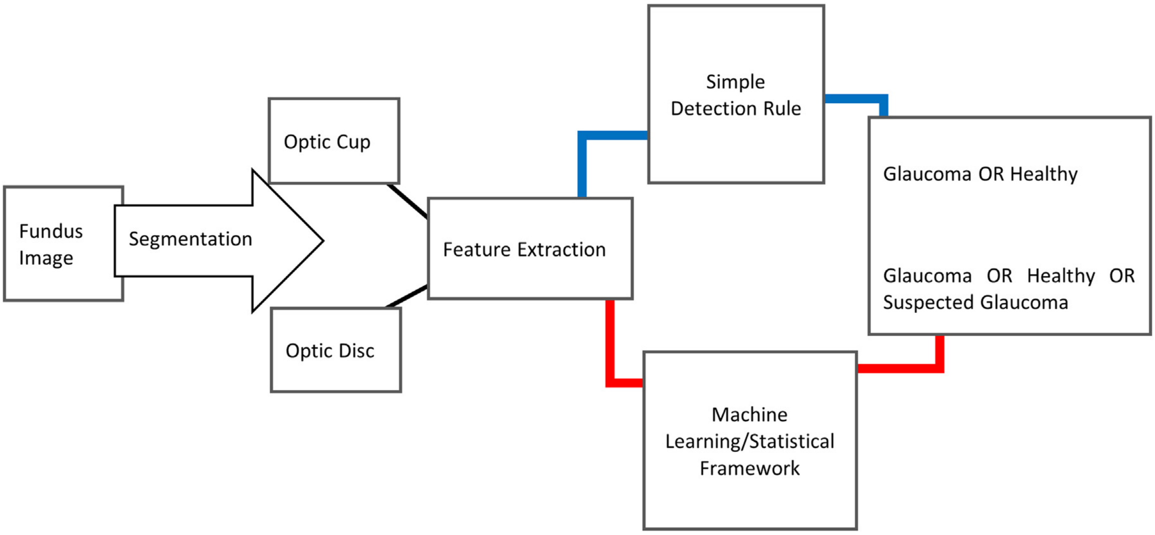

Fig. 4 – Pathways of frameworks for two-step AI-enabled glaucoma detection.

els/algorithms for glaucoma detection. In this review, 12 papers were identified as using the logical rule-based framework, while 24 papers used machine learning/statistical modeling frameworks.

5.3. *Logical rule-based AI frameworks for glaucoma detection from segmented images*

We use the term logical rule-based frameworks to refer to frameworks that use a set of simple IF-ELSE rules (Fig. 4). For such methods to work, the optic cup and disc are first segmented, then some clinically interpretable imaging features are obtained from the segmented image. Such clinically interpretable imaging features can include variations of the CDR (i.e., vCDR ratio, ACDR and RDR) and measurements from the NRR (i.e., NRR area, area in quadrants, ISNT rule compliance). These features are then used in the framework via IF-ELSE formats for glaucoma classification as presented in (Table 2). In the following text, we reflect on the key aspects of the reviewed papers that apply a logical rule-based AI framework.

Clinical features used by logical rule-based AI frameworks. The success of the logical rule-based frameworks is highly depen-

Table 2 – Details of the reviewed papers proposing logical rule-based AI frameworks. Afll papers considered two groups (glaucoma vs healthy), except for Issac and Dutta (2019) and (Soorya et al, 2018) whom both had three groups (healthy, glaucoma or suspected glaucoma): if the features obtained from the fundus image did not meet the criteria for glaucoma or healthy group, this was then classified as suspected. (- represents information not provided).

Paper	Features	Feature	Rule for Glaucoma Classification	Accuracy	Sensitivity	Specificity	Testing Data	Datasets
(Božić-Štulić et al., 2020)	1	ACDR	>0.3	96.8%	-	-	200	1
(Dutta et al., 2014)	1	vCDR	>0.75	90.0%	-	-	10	1
(Agarwal et al., 2015)	1	ACDR	>0.3	90.0%	-	-	20	1
(Ahmad et al., 2016)	1	vCDR	>0.5	92.0%	93.0%	88.0%	100	1
(Dutta et al., 2018)	1	ACDR	>0.26	83.0%	-	-	101	1
(Soorya et al., 2018)	1	vCDR	>0.7	97.0%	96.5%	98.0%	215	1
(Mvoulana et al., 2019)	1	ACDR	>0.63	98.0%	100.0%	94.4%	51	1
(Ong et al., 2020)	1	ACDR	>0.5	86% * BAC	82.0%	89.0%	133	1
(Das et al., 2016b)	2	vCDR ISNT	vCDR>0.5 AND ISNT violation	94.0%	92.6%	94.5%	244	5
(Issac and Dutta, 2019)	3	ACDR vCDR ISNT	ISNT rule violation AND vCDR > 0.6 OR ACDR < 0.25	93.0%	94.0%	96.0%	364	1
(Vijapur and Kunte, 2017)	3	ACDR RDR VRI	ACDR > 0.4 OR RDR < 0.6 OR VRI < 0.3	-	93% (Private Database) 87% (HRF)	92% (Private Database) 87% (HRF)	150 (Private Database) 30 (HRF)	2
(Neto et al., 2021)	1	ACDR	>= 0.2	-	82%	86%	660	3
	1	vCDR	>= 0.5	-	89%	79%		
	1	HCDR	>=0.5	-	82%	64%		

dent on the imaging features used. From the 12 papers identified, nine of the papers used one feature, one paper combined 2 features, and2 papers combined three features for their proposed detection rule (Table 2). The most frequently used feature was the ACDR, used by seven different frameworks. Following this, the vCDR was used by 6 frameworks, and a variation of the ISNT rule was exploited by two frameworks. The features of vessel ratio index (VRI) and RDR were both used once in combination with other features.

Logical rule-based AI frameworks using one feature. Variants of the CDR parameter have proven to be popular due to their clinical value, interpretability, and cheap computation from a segmented fundus image. However, some authors have criticized the use of a CDR feature alone, stating that the feature is a limited and incomplete parameter for classifying glaucomatous neuropathy (3,43,61).

The vCDR was used alone in a detection rule by Dutta and coworkers (32) with a reported accuracy of 90%. This framework was tested on a small sample of 10 images thus, only one image was incorrectly classified. The one incorrectly classified image displayed a vCDR of 0.6 which their rule classified as healthy yet the ground truth from ophthalmologists marked the observation as glaucomatous. Although a small study, this example highlights why using the vCDR alone can be problematic. Clinically, it is known that healthy individuals with a large disc can display large vCDR values, and conversely, glaucoma patients with a small disc can have small vCDR values^{L}. The authors also recognized this pitfall and propose that future work should consider incorporating other clinically interpretable features.

Three other reviewed papers considered the vCDR alone. Ahmad and coworkers obtained an accuracy of 92%, sensitivity of 93%, and specificity of 88% (6). While Sooryaand coworkers obtained an accuracy of 97%, a sensitivity of 96.5%, and specificity of 98% (60). Both frameworks (60,6) only tested their approach on a dataset acquired from one source which limits the conclusions that can be made about the frameworks' generalizability. Conversely, Neto and coworkers proposed 3 rules for glaucoma classification using the features of vCDR, hCDR and ACDR independently (70). They found the optimal results when using the vCDR, this gave a sensitivity of 89% and a specificity of 79%. Thus, although vCDR may be a limited parameter when used independently, it is better than the parameters of hCDR and ACDR in this case (70). Note that Neto et al tested their approach on a larger database of 660 images (Table 2).

Further work by Dutta and coworkers (31) proposed the use of the ACDR independently. The authors stated that the parameter of the ACDR is more appropriate than the vCDR parameter for glaucoma classification. They reasoned that the vCDR parameter assumes that the optic cup and disc are virtually circular; thus, the parameter will not account for any shape irregularities that occur with glaucoma neuropathy.

When using the ACDR alone, the reported accuracies from three papers ranged from 83% (31) to 90% (4) and 96.8% (14) (Table 2). Note that all three papers did not provide the metrics of sensitivity or specificity and used only one dataset. Two other papers (67,72) also used the ACDR parameter alone. Mvoulana et al' (67) analysis yielded an accuracy of 98%, sensitivity of 100% and specificity of 94% and Ong et al' analysis (72) yielded a balanced accuracy of 86% and a sensitivity and specificity of 82% and 89% respectively.

Logical rule-based frameworks using two or more features. Rather than using one feature alone, Das et al (25) proposed combining the vCDR with the ISNT rule for their detection rule. They classified an image as 'healthy' if the vCDR < 0.5 and it satisfies the ISNT rule, otherwise, the image was labeled as glaucomatous. Upon inspection of the framework's misclassifications, they determined that these occurred due to the segmentation step rather than the features used (25). Thus, highlighting the importance of accurate segmentation methods in the first step of the framework.

Vijapur and Kunte (95) used the 3 features of ACDR, rim-to-disc ratio, and vessel ratio index (Table 2). The authors cite that their detection rules were determined after consultations with ophthalmologists to ensure they were clinically relevant and appropriate (95). Their framework introduced the novel idea of segmenting blood vessels and accounting for this within glaucoma classification. However, further external testing is required to evaluate whether the vessel ratio index feature is generalizable to images from other sources.

Three clinically interpretable imaging features: vCDR ratio, ACDR & ISNT rule compliance were used by Issac and Dutta (46), the authors used a logical rule presented in a hierarchical IF-ELSE format (Table 2). This framework resulted in an accuracy of 93%, sensitivity of 94%, and specificity of 96% (46). In frameworks when rules are used in a hierarchical format such as this, it is important to note which features are first in the chain. While it is widely used in practice, the ISNT rule is shown to be less reliable than the vCDR parameter; thus, more errors could occur by applying the ISNT rule first (75).

Das and coworkers proposed the use of vertical cup-to-disc ratio in combination with the ISNT rule (25), the method was tested on four publicly available datasets and one private dataset. This framework resulted in an accuracy of 94%, sensitivity of 92.6%, and specificity of 94.5% (25). Following this, Issac and Dutta used the ACDR parameter with the vCDR parameter and the ISNT rule, yielding an accuracy of 93%, sensitivity of 94%, and specificity of 96% (46). Finally, the paper by Vijapur and Kunte used the ACDR with the RDR parameter and vessel ratio index (95). They obtained a sensitivity of 93% and specificity of 92%; the accuracy of the framework was not provided (95).

5.4. Machine learning/statistical modeling-based AI frameworks for glaucoma detection from the segmented image

The machine learning or statistical modeling–based AI frameworks differ from the logical rule-based AI frameworks as they implement a mathematically complex classifier to perform the classification of glaucoma. Alike to the logical rule-based AI frameworks, they can make use of clinically interpretable features extracted from a segmented fundus image, but different from the logical rule-based AI frameworks, they can also create and utilize abstract features and exploit these within machine learning or statistical modeling classifiers. The following section presents the findings of the 24 papers identified in this review that implement a machine learning or statistical modeling-based AI framework.

5.4.1. *Machine learning/statistical modeling-based AI classifiers and their reported performance*

The machine learning/statistical modeling frameworks differed from one another by the type of classifiers they implemented (Table 3). Support vector machines (SVM's) were the most popular classifier, being used in 11 out of 17 papers. The clustering methods of M-Mediods and K-nearest neighbours (K-NN) were used by one paper each and the ensemble classifiers of Random Forest (RF), dynamic ensembling and XG-Boost were all proposed once. Additionally, two papers used Linear Mixed Effects (LME) modeling. The remaining frameworks proposed a variant of a neural network (NN) for classification. Note that, Table 3 only presents the optimal classifier used in the frameworks. That is, many papers propose a range of classification models/algorithms and present the classifier which worked optimally. The type of optimal classier also depends on the dataset used.

Support Vector Machine Classifiers. A total of 11 studies used support vector machine classifiers, and two different kernel functions were selected within these. The radial basis function (RBF) kernel was used by five frameworks and the linear kernel was used by four frameworks. From the papers using the RBF kernel, Issac and coworkers (47) obtained an accuracy of 94%, sensitivity of 100%, and specificity of 90%. Krishnan and coworkers (55) only provided the F1 score as a metric, which was 91%. The framework proposed by Agarwal and coworkers obtained an accuracy of 90%, sensitivity of 100% and specificity of 80% (5); while the framework by Khalil and coworkers combined two support vector machines with RBF kernels and achieved an accuracy of 92.9%, sensitivity of 87.5% and specificity 90.84% (52). Khalil and coworkers found significant improvement in classification capabilities was achieved by combining the outputs of the support vector machine classifiers and considering a range of structural and textural features. A more recent study by Kang and coworkers resulted in an accuracy of 85.06%, sensitivity of 81.95% and specificity of 88.28% (49).

From the four papers that used support vector machine classifiers with linear kernels, Narasimhan and Vijayarekha only provided the metric of accuracy which was 95% (69). Mukherjee and coworkers obtained an accuracy of 87%, sensitivity of 86.4% and specificity of 90% (65). More recently, Pathan and coworkers achieved an accuracy of 96.66%, sensitivity of 100% and specificity of 95% with the publicly available DRISHTI database but on external testing with a private database, this reduced to an accuracy of 90%, sensitivity of 93.47% and specificity of 91.2% (59). Xu and coworkers proposed a linear kernel SVM in combination with a decision rule (96). Firstly, if RNFLD were present this was marked as glaucoma. If not, then the SVM was applied for the decision output. This novel method resulted in a sensitivity and specificity of 96.1% and 95.6% respectively. Furthermore, Xu and coworkers implemented external testing; this achieved the metrics of 98.4% sensitivity and 94.1% specificity; indicating the generalizability of their adopted approach. Deepika and Maheswari did not specify the kernel used, this framework yielded an accuracy of 91.67%, sensitivity of 90% and specificity of 93.3% (29). Likewise, Yunitasari and coworkers did not specify the kernel used; their proposed framework achieved an accuracy of 95%, sensitivity of 91.4% and specificity of 95.6% (97).

Clustering classifiers Clustering methods were used by two frameworks. The k-nearest neighbors algorithm (K-NN) was proposed by Lotankar and coworkers, achieving an accuracy of 99.2%, sensitivity of 86.7% and specificity of 84% (59). The framework of Akram and coworkers used a clustering method of M-Medoids (8). They proposed that there is variation in the number and distribution of the samples within the two classes (healthy & glaucomatous) and via employing multivariate m-modeling and classification, they could handle multimodal distribution of samples within the two classification groups (8). This method was tested on five datasets; the accuracy across the datasets ranged from 86.7% to 94.4%, sensitivity from 75% to 93.3% and specificity from 87.1% to 97.1% (8).

Random Forest classifier. A Random Forest classifier was proposed by Zahoor and Fraz (98). This method resulted in an accuracy of 95.3%, sensitivity of 96.31% and specificity of 95.33%. However, the authors state the use of the publicly available High-Resolution Fundus Image (HRF) database but removed nine of the total 36 fundus images without explanation.

XGBoost classifier. Afolabi and coworkers proposed an XG-Boost classifier resulting in an accuracy of 88.3% and AUC of 93.6% via 5-fold cross-validation (3).

Dynamic ensemble method. Zulfira and coworkers implemented a dynamic ensemble classifier, they used three publicly available datasets independently; the accuracy ranged from 90% to 91%, sensitivity from 86% to 90% and specificity from 86% to 89% (100). Their choice of a dynamic ensemble classifier was to handle the imbalanced datasets (i.e., different numbers of images for the three groups: healthy, mild glaucoma and severe glaucoma).

Linear mixed-effects statistical modeling. A linear mixed-effects (LME) modeling approach was used by two papers (61,54). This framework was originally proposed by MacCormick and coworkers and yielded an AUROC of 99.7%, sensitivity of 100% and specificity of 98.3% on internal testing. The proposed framework then employed external validation using the publicly available RIM-ONE V3 dataset, the AUROC obtained was 91% (61). A disadvantage of such an approach is in requiring the segmented image of healthy eyes to follow a statistical model with a plausible number of parameters. This is not always possible, however, in the case of glaucoma, this was a suitable approach. The authors determined that the contours of the optic cup and disc appeared to be two-centered ellipses in healthy eyes and additionally, they included a technique to account for each eye displaying different disc sizes– all of which were captured in the statistical model. Using this information, the classification of glaucoma was then based on a deviation of the contours from the model of healthy eyes.

This framework was then improved by Adithya and coworkers who incorporated further relevant parameters (ACDR and group variance) to improve the model performance. They achieved an AUC of 0.997 via internal testing and 0.969 on external testing (54).

Neural network classifiers. A multi-layer perceptron was proposed by Perdomo and coworkers with the final stage being

Table 3 – Details of the reviewed papers proposing machine learning/statistical modeling-based AI frameworks. Twenty papers considered binary classification (glaucoma vs healthy). Four papers (Khalil et al, 2017), (Perdomo et al, 2018), (Yunitasari et al, 2021) and (Zufira et al, 2021) proposed three classes (glaucoma, suspect glaucoma, and healthy). Krishnan et al, 2020 used only F1 score as a quality of classification metric, which was 91%.

Paper	Classifier	Features	Features	Data	Accuracy	Sensitivity	Specificity	AUROC	Validation
(Zahoor and Fraz, 2018)	Random Forest	10	Area of OC & OD, ACDR, Area of NRR, HCDR, vCDR, Area of ISNT regions (4)	RIM-ONE & HRF	95.3%	96.3%	95.3%	-	-
(Deepika and Maheswari, 2018)	SVM	4	ACDR & 3 statistical features from blood vessels	HRF	91.7%	90.0%	93.3%	-	60:40
(Issac et al., 2015)	SVM (RBF kernel)	3	ACDR, NRR Area & Blood Vessel Ratio	Private	94.0%	93.8%	94.0%	-	LOOCV
(Lotankar et al., 2015)	K-NN	4	vCDR, ACDR, RDAR & H-VCDR	Private	99.2%	86.7%	84.0%	-	10-Fold CV
(Pathan et al., 2021)	SVM (linear kernel)	10	ACDR, NRR Area, Colour (4) & Texture (4) features	DRISHTI	96.7%	100.0%	95.0%	-	10-Fold CV
				Private	90.0%	93.5%	91.2%	-	
(Kausu et al., 2018)	MLP	2	ACDR & Texture Feature (Energy)	Private	97.7%	98.0%	97.1%	-	10-Fold CV
(Krishnan et al., 2020)	SVM (RBF kernel)	1	vCDR	DRISHTI	-	-	-	-	50:50
(Agarwal et al., 2015)	SVM (RBF kernel)	2	ACDR & RDR	Private	90.0%	100.0%	80.0%	-	70:30
(Akram et al., 2015)	M-Mediods	10	vCDR & RDR Spatial Features (5) Spectral Features (3)	DRIVE	92.5%	83.3%	94.1%	-	70:30
				DIARETDB1	94.4%	75.0%	96.3%	-	
				DRIONS-DB	93.6%	86.7%	94.7%	-	
				HEI MED	86.7%	84.2%	87.1%	-	
				MESSIDOR	89.0%	84.0%	94.4%	-	
				HRF	91.1%	93.3%	90.0%	-	
				GlaucomaDB	90.8%	85.7%	92.9%	-	
(MacCormick et al., 2019)	LME	24	pCDR (24 CDR's)	ORIGA	-	96.6%	99%	99.7%	70:30 & 100 bootstrapped samples
				RIM-ONE	-	-	-	91.0%	External Validation
(Narasimhan and Vijayarekha, 2011)	SVM (linear kernel)	2	ACDR & ISNT Ratio	Private	95.0%	-	-	-	70:30
(Mukherjee et al., 2019)	SVM (linear kernel)	8	vCDR, ACDR, dCDR, notch factor, S&I Distance, ISNT rule.	Private	87.0%	86.4%	90.0%	-	5-Fold CV
(Karkuzhali and Manimegalai, 2017)	FFBPNN	3	vCDR, ISNT Ratio & DOO	DRISHTI	100.0%	100.0%	100.0%	-	50:50
(Kang et al., 2020)	SVM (RBF kernel)	8	vCDR, ISNT score, length, area, distance from OD	Private	85.1%	82.0%	88.3%	-	60:40

(continued on next page)

Table 3 (*continued*)

Paper	Classifier	Features	Features	Data	Accuracy	Sensitivity	Specificity	AUROC	Validation
(Khalil et al., 2017)	SVM (RBF kernel)	62	vCDR, RDR, Cup shape & texture/intensity features	GlaucomaDB	94%	96%	92%	-	10-Fold CV
(Raja and Ramanan, 2019)	DLRNL	6	ACDR, NRR Area, BVR & Texture features	HRF	89.0%	-	-	-	-
(Perdomo et al., 2018)	MLP	19	Geometric (2), Ratio (7), Distances (4) & Axis (4) features	RIM-ONE & DR	89.3%	89.5%	88.9%	82.0%	70:30
(Zufira et al., 2021)	DES-MI (Dynamic Ensemble Method)	7	6 Features from GLCM (contrast, dissimilarity, homogeneity, energy, correlation, angular second moment) & ACDR.	RIM-ONE	91%	86%	87%	-	5-fold CV
				KAGGLE	90%	90%	86%	-	
				MESSIDOR	91%	90%	89%	-	
(Xu et al., 2021)	Simple rule on RNFLD then SVM (linear kernel)	3	RNFLD presence, MCDR (mean cup to disc ratio) and ISNT score	Private	-	96.1%	95.6%	98.1%	80:20
				Private		98.4%	94.1%	98.3%	External Testing
(Mansour et al., 2021)	Perceptron based Convolutional Multilayer Neural Network	2	vCDR & Holistic Features	DRISHTI	-	-	-	97.1%	-
(Yunitasari et al., 2021)	SVM	7	vCDR, optical disc area, optical cup area, optical disc perimeter, optical cup perimeter, optical disc circularity and optical cup circularity.	Private and DRISHTI	95%	91.37%	95.86%	-	50:50

(*continued on next page*)

Table 3 (*continued*)

Paper	Classifier	Features	Features	Data	Accuracy	Sensitivity	Specificity	AUROC	Validation
(Singh et al., 2021)	MLP	20	Homogeneity Contrast Correlation Standard deviation disc Mean disc Entropy disc Energy disc Standard deviation cup Mean cup Entropy cup Energy cup Radius disc Area disc Radius cup Area cup Cup-to-disc ratio Inferior region area Superior region area Nasal region area Temporal region area	DRIONS	95.82%	98.59%	98.6%	-	70:30
(Adithya et al., 2021)	Linear Mixed Effects Model	27	pCDR (24 CDR's), ACDR & 2 variance parameters	ORIGA	0.989	0.974	-	0.997	50:50
				DHRISTI	0.947	0.923	-	0.969	External Testing
(Afolabi et al., 2021)	XGB (Extreme Gradient Boost)	10	CDR at 10 locations	RIMONE V3, DRISHTI GS	88.3%	-	-	93.6%	5-Fold CV

composed of two fully connected layers with 64 hidden and 3 output units (74). The batch size, the number of epochs and optimal features were determined via a grid search. The binary classification achieved an accuracy of 89.3%, sensitivity of 89.5%, specificity of 88.9% and AUC 82%. Using multi-class classification (healthy, suspected glaucoma, and glaucoma), they provided the metrics of precision and recall which were 0.76 and 0.72 respectively.

Raja and Ramanan proposed the use of Damped Least-Squares Recurrent Deep Neural Learning Classification (DL-RNL) (77). The classification was performed on the output layer using soft sign activation functions resulting in an accuracy of 89% however, no other performance metrics were specified. The paper by Karkuzhali and Manimegalai (50) considered a range of networks, the best performance was found when using the Feed Forward Back Propagation Neural Network (FFBPNN) and the Distributed Time Delay Neural Network (DTDNN); each of these provided an accuracy of 100% and sensitivity and specificity of 100%. Note that they tested on a small subsection of the publicly available DRISHTI dataset consisting of just 26 images. Kausu and coworkersused a multi-layer perceptron and obtained an accuracy of 97.67%, sensitivity of 98% and specificity of 97.1% (51). Note they did not provide any detail of the multi-layer perceptron (i.e., number of neurons in each layer, hyperparameter tuning etc.).

More recently, Singh et al proposed an MLP using twenty clinical features (Table 3); on internal testing, this resulted in an accuracy of 95.8% (82). They found that the MLP provided higher performance metrics than other popular machine learning classifiers (*K.N.N., S.V.M.* and Naïve Bayes). Mansour and coworkersproposed a perceptron-based convolutional multilayer neural network, the performance metric of AUC was 97.1% on internal testing (62).

5.4.2. *The machine learning/statistical modeling AI frameworks utilize clinically interpretable image features as well as abstract image features*

The machine learning/statistical modeling-based AI frameworks reviewed used clinically interpretable and abstract image features (Table 3). Across all the papers reviewed, each framework used some variant of the CDR parameter, highlighting the significance of the parameter in glaucoma classification. Thirteen of the papers used clinically interpretable imaging features (i.e., vCDR ratio, NRR area etc.), 11 papers proposed the use of novel spatial/spectral/texture/color features.

The use of spatial features by Akram and coworkerss(8) was motivated by the fact that the area of the optic cup changes from the normal to the glaucomatous eye. In keeping with MacCormick and coworkers(61) and Adithya and coworkers (54), they state the use of the vCDR parameter alone was limited due to glaucoma manifesting at any direction in the ONH. Whereas Akram and coworkers combined the RDR parameter with spatial and spectral features (8), MacCormick and coworkers and Adithya et al proposed using a profile CDR (pCDR) which quantifies the optic nerve rim consistency around the whole disc at 15-degree intervals (54,61). Moreover, due to the use of linear mixed-effects modeling by MacCormick and coworkersand Adithya and coworkers, random effects were incorporated to indirectly take account of the size of the optic disc. The difference between the original work by MacCormick and coworkers and Adithya and coworkersis the use of the ACDR feature by Adithya and coworkers, and the inclusion of variance parameters to better capture the difference between the healthy and glaucoma group (54).

A similar approach was adopted by Afolabi and coworkers (3). Their framework used the CDR measured at 10 locations around the ONH citing their framework eliminates the challenge of selecting the CDR threshold as required in logical rule frameworks (section 5.3). Yet, in contrast to previous works, they state that 10 CDR values are optimal as 5 did not give a full view of the changing geometry of the optic cup and disc and extracting more than ten 10 features only resulted in duplication of data (3). However, their approach is yet to be tested on external data.

The framework by Kausu and coworkers (51) exploited clinically interpretable imaging features and abstract features in combination. Wavelet features were considered as the authors argued that texture features alone are not enough, as they do not consider frequency information. Yet, by exploiting the wavelet transform, frequency and spatial information would be considered. Kausu and coworkers (51) used the minimum redundance maximum relevance (mRMR) feature selection technique to determine the optimal features from the collection of clinically interpretable and wavelet features. However, in the end, the best performance was obtained when only using two features: vCDR and the second-order texture feature – energy. While the vCDR parameter is clinically interpretable, the second-order feature of energy is an abstract feature.

Similar features were exploited by Singh and coworkers(82) and Zulfira and coworkers(100), both used a combination of clinical features (i.e., CDR) and abstract features. Both made use of Gray-Level Co-occurrence matrix (GLCM) features (i.e., contrast, energy, etc.) (Table 3). Regarding clinical features, Zulfira and coworkers used ACDR and accounted for PPA via GLCLM features (100). While Singh and coworkers highlighted the importance of the ISNT rule and incorporated features to account for this (e.g., inferior/superior area) (82).

Correlation-based feature selection was applied by Pathan and coworkers73). They began with 54 color features, 12 texture features, and 2 clinically interpretable features. Following feature selection, 10 features (2 clinical, 4 color & 4 texture) were deemed relevant and applied in the final framework.

Mukherjee and coworkers(65) proposed a framework using eight features (Table 3). They compared this framework with the same methodology but using only the CDR feature yet, they found this resulted in significantly lower performance metrics. Thus, indicating the relevance of the other parameters used; however, this is to be further examined to test the generalizability of the other features for glaucoma classification with external datasets (65). Similarly, Khalil and coworkers(52) found an improved performance by combining structural and textural features for classification (Table 3).

5.5. *Approaches to segmentation*

Intuitively, the success of a multi-step framework depends on the type and success of the automated ONH segmentation used in the first step of the framework. Although the focus of this review is not to assess the automated segmentation methods, in this section we give a brief overview of the approaches to segmentation used. Briefly, some automated segmentation methods focus on color intensity and texture-based thresholding. Some advanced methods employ fully convolutional neural networks. The point is that there are many different approaches to segmentation with differing degrees of success. In the segmentation of the ONH, it is well-known that the optic cup is much more challenging to segment than the optic disc due to the low contrast between the optic cup and the neighboring disc region (31). As such, there are very few papers focused on developing optic cup segmentation methods.

5.6. *Glaucoma disease groups*

From the 34 papers identified in this review, 28 (82%) performed binary classification (healthy or glaucoma), and 6 performed multi-class classification. Across the papers performing multi-class classification, 4 classified images by healthy, suspected glaucoma, or glaucoma, and their method for incorporating the suspected class differed. Moreover, 2 proposed multi-class classification; however, they differentiated between the classes via severity (e.g., healthy, mild glaucoma, and severe glaucoma).

The framework of Khalil and coworkers (52) used a combination of clinically interpretable and abstract features in two support vector machine classifiers (one support vector machine using structural features and one support vector machine using textural features) for glaucoma classification. They proposed that, if the outputs of the 2 support vector machine classifiers did not agree (i.e., one support vector machine provides the outcome healthy and the other glaucomatous), they would classify this image as ‘suspect glaucoma’.

Perdomo and coworkers(74) proposed a multi-layered perceptron with 3 output units using 19 morphometric features. They used the publicly available RIM-ONE V3 dataset which comprises 35 suspected glaucoma fundus images for training/testing their framework to handle the suspected class. Although they showed high performance metrics on binary classification, the performance on multi-class classification was inferior; the metrics of precision and recall were 0.76 and 0.72 respectively (59). Thus, their framework was not optimal when considering the suspected glaucoma class.

More recently, frameworks by Soorya and coworkers (60) and Issac and Dutta (46) applied logical rule-based AI frameworks with thresholds for glaucomatous and healthy; if the features obtained from the segmented fundus image did not meet the criteria for the glaucoma or healthy group, this was classified as suspected glaucoma (Table 2).

Zulfira and coworkers proposed a framework to provide glaucoma severity: healthy, mild glaucoma or severe glaucoma (100). To achieve this, they used a dynamic ensemble classifier and features to represent PPA and CDR (Table 3). They found the highest accuracy in images with severe glaucoma but noted a lower accuracy in the mild glaucoma images as they were frequently misclassified as healthy. Thus, highlighting the difficulty in distinguishing between the subtle differences that mark an eye as healthy or mild glaucoma. Moreover, they do not provide information regarding the ground truth criterion, or the number of experts used. It would be of interest to know how the ‘mild’ glaucoma group is defined. Although their proposed method outperformed the deep learning-based U-net when evaluated using the same datasets, it has a drawback in requiring manual ONH detection for all images by an expert (100).

A similar approach was adopted by Yunitasari and coworkers; however, they categorized images as early, moderate, and advanced glaucoma (97). Using an SVM and a combination of clinical features (Table 3) they tested their approach on 40 images and found encouraging results highlighting that automatic glaucoma severity marking could be a possibility. Yet, no information is provided regarding the ground truth definitions (i.e., the difference between the early and moderate glaucoma groups). Furthermore, the clinical application is to be considered, for example, how would glaucoma severity aid clinicians in practice and/or how is the framework going to work with healthy images as these have not been considered to date.

5.7. *Approaches to validation of methods and the reporting of performance metrics*

The approach to validation in logical rule-based AI frameworks. In the papers that used a logical rule-based AI framework, the approach to validation differed as they have no training component within their frameworks. The only means of validation per se (using a logical rule-based framework) is to acquire datasets from a range of sources to evaluate if their proposed rules are generalizable/appropriate. Of the 12 logical rule-based frameworks (Table 2), 9 (81%) used one dataset, 1 paper used 2 datasets, 1 paper used 3 datasets, and 1 paper used 5 datasets. As such, the majority of papers using logical rule-based frameworks did not consider validation of their proposed frameworks.

Considering the performance metrics presented, 6 papers (50%) presented the performance metrics of accuracy, sensitivity & specificity while the remaining 6 papers did not. Four of the papers only provided their accuracy result, and 2 papers did not provide accuracy, only sensitivity & specificity.

The approach to validation in machine learning/statistical modeling AI frameworks. The machine learning/statistical papers differed in their approach to framework validation. The approach of 10-fold cross-validation was used by 5 papers, 5-fold cross-validation was used 3 times, and leave-one-out cross-validation was also used once. The remaining papers used a data split for validation. A 70:30 split was used 4 times whilst a 50:50 split was used 5 times. External validation was only used by three papers. Seven papers used more than one database within their frameworks. For this, they either trained or tested their model individually on the different databases or they combined the databases and then trained and tested on the data (Table 3).

In addition to conducting internal/external validation, some of the reviewed papers compare their AI method with previously published methods. Fourteen of the 24 machine learning papers compared their proposed methodology with previously published methods while 17 papers compared their methods with at least one other method proposed by themselves.

Table 4 – Databases used for development (training and testing) of the reviewed two-step AI-enabled glaucoma detection frameworks.

Database	Number of Times Used	Total Number of Images	Healthy	Glaucoma	Suspected	Annotators
HRF	7	45	15	1	NA	-
RIM-ONE V1	1	169	118	51	NA	5
DRISHTI	12	101	70	31	NA	4
Messidor	4	100	72	28	NA	-
Drions	3	110	95	15	NA	-
DiaretDB	2	89	81	8	NA	4
RIM-ONE V2	1	455	255	200	NA	1
RIM-ONE V3	5	159	85	39	35	2
GlaucomaDB	2	120	85	35	NA	-
DRIVE	1	40	34	6	NA	2
HEI MED	1	50	31	19	NA	2
ORIGA	2	650	482	168	NA	-
Private	17	NA	NA	NA	NA	NA

Considering the performance metrics reported, 17 papers disclosed metrics of accuracy, sensitivity and specificity. Only 4 papers presented metrics for AUC, and 1 paper presented no metrics other than the F1 score. Additionally, 2 papers only presented the accuracy metric results.

5.8. *Databases used for development and testing of the AI frameworks reviewed*

Within the frameworks highlighted in this review, a range of publicly available and private databases were used, an overview is provided in Table 4.

5.8.1. *Publicly available datasets*

DRISHTI dataset. From the papers identified, the DRISHTI database (84) was the most popular database being used 12 times. The database comprises 101 fundus images (31 healthy and 70 glaucoma) acquired at Aravind Eye Hospital, Madurai, India. This dataset is of a single population as collected images are from subjects who are Indians. The images were taken with the eyes dilated using the following data collection protocol: centered on the optic disc with a field-of-view of 30-degrees and dimension 2896 × 1944 pixels. Low-quality images (poor contrast, positioning of optic disc region, etc.) were not used. The ground truth for the region boundaries, segmentation soft maps and CDRs by 4 different ophthalmologist experts (with varying clinical experience) is provided. The database is split into 50:51 training: testing. Note that, to access the ground truth for the test data, a researcher must register with the data owners (83).

High-Resolution Fundus (HRF) dataset. The HRF dataset (15) was used by seven of the reviewed papers. In comparison to the other databases available, this database is small, comprising 45 fundus images in total. The images were collected at the same clinic in the Czech Republic (71). Of the 45 images, 15 are glaucomatous, 15 healthy and 15 are labeled as diabetic retinopathy. The database is publicly available and in an easily downloadable format online. All fundus images were acquired with a mydriatic fundus camera CANON CF–60 Uvi equipped with a CANON EOS–20D digital camera with a 60-degree field of view (FOV). The image size is 3504 × 2336 pixels (56). The database curators do not state how many ophthalmologists were used for the ground truth. As well as the fundus images, researchers can access the Field Of View (FOV) masks, vessel segmentation, and optic disc gold standards provided by 3 experts (71). Whether the images were obtained in a dilated state is not disclosed.

Messidor dataset. The Messidor database (28) was used in four reviewed papers. It contains a total of 1200 images of different diseases, but only 100 images are annotated for glaucoma. Of the 100 fundus images, 28 are glaucomatous and 72 are healthy. The images were acquired by 3 ophthalmologic departments in France using a color video 3CCD camera mounted on a Topcon TRC NW6 nonmydriatic retinography with a 45-degree field of view. To access the dataset, the researcher is required to submit a form that is evaluated by the data owners, and they decide upon the validity of the request and provide permission (28).

ORIGA dataset. The ORIGA database (99) was used in 2 reviewed papers. The ORIGA database consists of 650 fundus images in total, 168 glaucomatous images and 482 randomly selected non-glaucoma images. The authors state that there are 336 images from the left eye and 314 from the right. The ORIGA database was formed using retinal image data collected from the Singapore Malay Eye Study (SiMES) (34) conducted by the Singapore Eye Research Institute. Each image is tagged with grading information (CDR, ISNT rule compliance, RNFL defects, notches, and PPA) and the manually segmented result of the optic cup and disc (99). Although it is stated that it is publicly available, it is not easily accessible from searching online. Moreover, no details are provided regarding the imaging device used (53).

RIM-ONE dataset. Four reviewed papers utilized the RIM-ONE databases (12,34,36). There are three different versions of the RIM-ONE databases: V1 and V2 which were used once and V3 – which was used five times. RIM-ONE V1 (34) was published in 2011; the dataset comprises 169 fundus images from different subjects. There are 5 groups: Normal, Early Glaucoma, Moderate Glaucoma, Deep Glaucoma, and OHT (Ocular Hypertension) which have 118, 12, 14, 14, and 11 images respectively. The RIM-ONE V1 database consists of 5 manual reference segmentations per image. This enables the creation of reliable gold standards, thus decreasing the variability among expert segmentations and the development of highly accurate segmentation algorithms (34). The fundus images were acquired from three different hospitals located in different Spanish regions (Hospital Universitario de Canarias, Hospital Clińico San Carlos and Hospital Universitario Miguel Servet). The authors of RIM-ONE state that compiling images from different medical sources guarantees the acquisition of a representative and heterogeneous image set (34). The images were captured using a Nidek AFC-210 nonmydriatic fundus camera with a 21.1-megapixel Canon EOS 5D Mark II body, with a vertical and horizontal field of view of 45°.

The RIM-ONE V2 dataset (35) comprises 455 fundus images (200 glaucomatous and 255 healthy), the ground truth for the images were provided by one expert ophthalmologist. The most recent version of the database is the RIM-ONE V3 which includes 159 fundus images with 85 healthy, 39 glaucoma and 35 suspected glaucoma. The images were taken by a nonmydriatic Kowa WX 3D stereo fundus camera (2144 × 1424 pixels) and 34-degree POV. The images were acquired at the Hospital Universitario de Canarias, and the ground truths provided by two experts (12).

GlaucomaDB dataset. The GlaucomaDB (52) database was used twice by frameworks in this review. The database is a subset of 120 fundus images from a larger database of 462 images gathered in a local hospital. The region/country of the local hospital was not disclosed. The images were captured using a TopCon TRC 50EX camera with a resolution of 1504 × 1000. The 120 images consist of 85 healthy and 35 glaucomatous, with the ground truths provided by 2 ophthalmologists (52). To access the database for research purposes, permission from the authors must be requested.

HEI MED dataset. The HEI Med Dataset (Hamilton Eye Institute Macular Edema Dataset) (39) is a collection of 169 fundus images, however, only 50 images are annotated for glaucoma detection. The HEI MED database was used by one framework. Of the 50 images, 30 are healthy and 19 are glaucomatous. The fundus images were collected at the Hamilton Eye Institute, United States of America, via a Visucam PRO fundus Camera (Zeiss) (53) and annotated by one ophthalmologist. The data is available on GitHub for public use.

DRIONS dataset. The DRIONS database was used three times by papers in this review. The database comprises 110 fundus images (95 healthy and 15 glaucomatous). The images were collected at the Ophthalmology Service at Miguel Servet Hospital, Saragossa, Spain. Images were removed if any form of cataracts were present. All images were obtained from subjects of Caucasian ethnicity. The images were acquired with a color analogical fundus camera, approximately centered on the ONH and they were stored in slide format. To have the images in digital format, they were digitized using a HP-PhotoSmart-S20 high-resolution scanner, RGB format, resolution 600×400 and 8 bits/pixel (18). The dataset is easily accessible and is available to download online.

DIARETDB dataset. The DIARETDB^{M} database consists of 89 color fundus images and was primarily developed for aiding diabetic retinopathy research; however, the database has been made publicly available and it has been assessed for glaucoma. The 89 fundus images are split into 81 healthy and 8 glaucomatous (8), and 4 medical experts were collected for the ground truth annotations. All images were captured using the same 50-degree field-of-view digital fundus camera with varying imaging settings at Kuopio University Hospital, Finland. The database is easily accessible for download online.

DRIVE dataset. The DRIVE database was used by one paper in this review. The database comprises 40 fundus images (34 healthy and 6 glaucomatous) annotated by 2 ophthalmologists (52). The images were acquired using a Canon CR5 nonmydriatic 3CCD camera with a 45-degree field of view (FOV). Each image was captured using 8 bits per color plane at 768 by 584 pixels. Although stated that the database is publicly available, it is not easily accessible online.

5.8.2. *Private datasets*

Private databases were popular in the development (training and testing) of the frameworks reviewed, a total of 17 private databases were used; however, as these are not publicly available, it was difficult to access detailed information on the databases if not provided directly by the authors. Many papers did not provide basic information other than the dataset size. Without all information regarding the datasets (i.e., patient cohort, imaging device, etc), it is difficult to draw conclusions regarding the robustness and generalizability of the proposed frameworks (as this is dependent upon the dataset used).

6. Discussion and conclusions

We present the first review, to our knowledge, of AI frameworks for glaucoma detection that utilize fundus images and produce ONH segmentation as the first step. By segmentation, we refer to a process of an image being automatically partitioned into 3 areas: optic cup, optic disc and neuroretinal rim. We identified 36 papers published between January, 2011 and December, 2021.

We focused on fundus imaging as it is the simplest modality of ONH assessment. The quality of fundus images may be sufficient for evaluating ocular health for the presence of glaucomatous neuropathy and due to its relatively low cost, fundus cameras are readily available in a range of settings. As such, there is the potential to exploit fundus images via AI to develop automatic glaucoma screening provisions, even for economically weak areas of the world. AI-supported color fundus images can help in two scenarios. Firstly, in a non-portable office-based environment–e.g., high-street optometry – it can assist in diagnosis and highlight patients for referral. This could reduce unnecessary referrals and thus reduce the burden on the health care sector. Secondly, it can be a part of portable devices in less well-resourced environments for use by an ophthalmic technician or nurse to screen. In both scenarios, the AI certainty element will be an important factor to consider for patient safety. Such screening provisions can have a clinical oversight to monitor the quality of the screening.

6.1. *Three key findings of this review*

6.1.1. *Glaucoma detection via two-step AI frameworks using fundus images present encouraging results*

We found that the two-step AI frameworks have presented promising results in their first step when identifying the contours of the optic cup and disc (i.e., segmentation of the ONH). We then identified two approaches to using features derived from the segmented fundus images: logical rule-based frameworks and machine learning/statistical modeling frameworks.

This review highlights that the glaucoma detection performance of the logical rule-based AI frameworks is limited due to the nature of using a set of rules (even more so when the rules have been derived from small homogenous datasets). We found that ten papers split one dataset for training and testing and reported accuracy ranged from 83% to 97%. Since this accuracy was determined via internal validation and on small datasets it must be interpreted with caution. One paper (Vijapur and Kunte (95)) did perform external testing (i.e., they used more than one data source). They reported two combinations of sensitivity vs specificity: 93 vs 92%, and 87 vs 87%. Across all papers, we found that there was no consensus on thresholds applied within the rules for glaucoma classification. That is, although many papers highlighted that their rule was based upon clinical relevance (as they were using a clinical parameter within their rule i.e., vCDR), the threshold used for the clinical parameter changed from one paper to another. Consequently, this highlights that a given threshold may only be appropriate for the dataset at hand. Moreover, as the majority of the logical rule-based AI frameworks did not implement any external testing, we are limited in understanding how the framework would work in screening strategies with data collected from different sources.

Regarding the machine learning/statistical modeling-based AI frameworks reviewed, we found that the reported accuracy was between 85.1% and 100%, predominantly reported via internal validation. The reported performance of some of the frameworks was comparable to that of the one-step approaches using DL. One of the current DL approaches is by Li et al (57) who reported an accuracy of 0.986. However more direct accuracy comparisons are required on the same testing datasets to give a fair comparison of the approaches.

6.1.2. *There is active research into developing two-step AI frameworks for glaucoma, where the first step is automatic detection of optic cup and disc contours*

We conducted this review by focusing on two-step AI frameworks that produce ONH segmentation as a first step. One key reason for this is the interpretability and explainability benefits that can be found when using segmented images within AI frameworks. It is known that the segmented contours of optic cup and disc can explain to the clinician why the AI has classified a given fundus image as glaucomatous or not. The two-step solution helps visually explain intermediate steps between the raw image and diagnosis (27). This can significantly aid in the development of trust within the AI and consequently the adoption of such methods within the practice of glaucoma detection. Moreover, such explainable AI methods can act as a support decision system such that the AI, clinician and patient can work together to decide upon treatment options and next steps.

6.1.3. *Color fundus images are actively studied for their potential use within AI-enabled glaucoma detection*

This review solely focuses on glaucoma detection frameworks using fundus imaging technology. This choice was guided by the fact that the detection of glaucoma in clinical practice is highly influenced by optic nerve head assessment via fundus imaging and the use of color fundus images is part of the guidelines for glaucoma diagnosis. Moreover, color fundus images are advantageous due to their lower cost in comparison with other imaging modalities and the technology is continuously developing such that they can consistently provide high-quality images capable of distinguishing glaucomatous neuropathy.

It should be highlighted that there is a distinction between large fundus cameras, costing many thousands of pounds or dollars, and the recently developed smaller mobile phone cameras that enable fundus imaging of the ONH at a considerably lower cost. While other imaging modalities such as OCT can provide additional information and are becoming more widely available, it is currently hard to see if lower-cost mobile OCT is possible and hence whether it will be available to less developed countries and remote areas. Yet portable fundus cameras are becoming increasingly accessible and viable, even within economically less fortunate countries.

6.2. *Three key unresolved issues of current knowledge and potential areas for future studies*

6.2.1. *There is a need to work on AI frameworks that utilize color fundus images and that provide contours of the optic cup and disc in their first step*

A direct comparison of all approaches for AI-enabled glaucoma detection methods is required. One-step AI approaches (end-to-end approaches, based on DL) need to be compared to two-step approaches reviewed here, on the same datasets. This will ensure a direct comparison can be made and one can consequently identify the benefits and drawbacks of both approaches. For now, we can identify that the advantage of the one-step AI is that it does not require such a large effort in terms of segmenting the fundus images and deriving clinical features from the segmentations; this is due to the nature and complexity of DL; however, a major disadvantage is in the lost interpretability (due to the black-box nature of DL) and in needing many annotated images to be trained. Such algorithms need to be studied together with two-step algorithms, to understand better which are more suited for glaucoma detection.

Furthermore, more research needs to be done on comparing imaging modalities. Specifically, investigations need to be made between OCT and fundus imaging to comprehensively compare both modalities such that we can determine which is most suitable for AI-enabled glaucoma detection frameworks.

Moreover, further research should be conducted regarding the development of AI-enabled glaucoma detection frameworks. For example, an area to be studied is the quantification of uncertainty of the outputs provided by the AI framework[J].

Also, the inclusion of other data sources needs to be investigated, e.g., patients' de-identified personal data, genetics data, visual fields data. This will simulate the clinical workflow as well as potentially improve the performance of the AI frameworks and help to explain AI outputs.

6.2.2. *There is a need to keep building and sharing suitable datasets*

There exist several large landmark clinical study datasets which were not used in the publications reviewed - despite being a very rich resource of clinical images for glaucoma diagnosis. This includes the United Kingdom Glaucoma Treatment Study (UKGTS) (37), the Ocular Hypertension Treatment Study (OHTS) (40), and the Northern Ireland Cohort Longitudinal Study of Ageing (NICOLA) (63). There are several possible reasons for the exclusion of such datasets. These datasets lack pixel-level image annotation of the optic cup and disc, which is required to train and validate segmentation models. Acquiring such annotations is a time-consuming task requiring collaboration between domain expertise and technical expertise. Furthermore, access to these datasets requires an application, payment and submission of a suitable protocol, which can act as barriers.

Moving on, our review has highlighted the increasing need for datasets to include the whole spectrum of glaucoma severities (not just glaucoma and normal, but also for glaucoma suspects). This is crucial to the development of AI frameworks that are useful in clinical practice as 'suspect glaucoma' is a case regularly observed by clinicians. Additionally, it is very important to collect and develop resource-rich longitudinal datasets such that disease onset and progression can be examined and incorporated into AI frameworks.

We also highlight that it is essential that sufficient details are provided alongside datasets. This includes the number of patients, the number of images acquired from each patient, whether both eyes are used (i.e., an image per eye) etc. All this information is important and relevant for researchers developing AI frameworks as such methods can be based upon hard assumptions. If these assumptions are not upheld due to the lack of information provided with the datasets, this can cause significant issues. Additionally, other information that should be recorded including the type of camera used, number of ophthalmologists for annotation of images/providing of ground truth, source of data, inclusion/exclusion criteria for data collection, etc., was also limited. In particular, when data has been acquired from a private source, there has been a scarce amount of information provided. A detailed description of the dataset used is critical for the assessment of the quality, reliability, suitability to produce the desired output, potential generalizability of any findings, and especially reproducibility of the methods (88).

Another important point to highlight is that further effort is required to ensure datasets are provided with suitable gold standards (aka ground truth). High-quality gold standards are crucial for AI development. A means of achieving this is acquiring annotations from multiple human graders. The reasoning for this is that ONH annotation (via fundus images) can suffer from large amounts of inter-observer variability–it is a subjective task (48,21). Using only one grader introduces bias into the ground truths upon which the AI is developed. A useful measure of the reliability of ground truth labels is an interobserver agreement between the labelers. By detailing the interobserver agreement, readers can make a judgment on the likelihood that the ground truth label is correct. This review has highlighted that only 3 of the 12 publicly available databases have more than two annotators. Whilst it should be standard to have more than 2 annotators, it should be recognized that obtaining manual annotation of images is not an easy task as it is time-consuming, expensive, and requires expertise.

Moreover, further work is required to improve the diversity of datasets. The use of the term diversity here refers to having fundus images captured from various devices, involving different patient ethnicities, and images taken in different lighting, contrast, and noise (12). The frameworks reviewed here are developed on datasets predominately acquired from one source and as such lack this diversity. A potential limitation of this is whether the quoted sensitivities and specificities will be generalizable to real-world patient cohorts where a range of factors can negatively impact the quality of the fundus images (66). Moreover, selection bias can be present if the dataset has been collected from homogeneous sources (i.e., using one ethnicity and/or specific hardware/imaging settings). Methods developed on such datasets are prone to generalization problems as one population data might have different characteristics that introduce bias in the proposed framework (80).

6.2.3. *There is a need to keep developing guidance for high-quality reporting of AI frameworks and promote following the guidance*

This review highlights that several publications lacked high-quality reporting–both in terms of datasets used and their glaucoma classification methodology. Some of the reviewed papers lacked the technical details regarding their classifier whilst others only provided a brief explanation of the methods selected. Often lacking sufficient detail was the model structure (i.e., hyperparameters used and their tuning mechanism). The limitation of not providing sufficient details of methods (i.e., technical AI details) is that this renders the paper unreproducible, a key criticism in the AI field.

There is a need to support the work of the initiative EQUATOR[K] which is a collaboration between experts in statistics, machine learning and computing – but it also involves specialized clinicians and policymakers. This initiative develops and provides detailed guidance for reporting, with a specific focus on guidance for medical studies involving AI.

7. Method of literature search

We used four databases to search for relevant literature: PubMed, Scopus, Web of Science and Medline (OVID). The search covered January 2011 until the end of 2021. The search strategies are detailed in supplementary file 1.

7.1. *Search terms*

7.1.1. *Database: Scopus*

(TITLE-ABS-KEY (glaucoma) AND TITLE-ABS-KEY (fundus OR retinal) AND TITLE-ABS-KEY (classif* OR discrim* OR diagnos*) AND TITLE-ABS-KEY (photograph* OR imag*) AND TITLE-ABS-KEY (“auto* detect*” OR “detect” OR “predict*”) AND TITLE-ABS-KEY (segment*))

7.1.2. *Database: PubMed*

(((((glaucoma[Text Word]) AND (fundus[Text Word] OR retinal[Text Word])) AND (classif*[Text Word] OR discrim*[Text Word] OR diagnos*[Text Word])) AND (photograph*[Text Word] OR imag*[Text Word])) AND (“auto* detect*”[Text Word] OR detect*[Text Word] OR predict*[Text Word])) AND (segment*[Text Word])

7.1.3. *Database: Web of Science*

TS = (glaucoma AND (fundus OR retinal) AND (classif* OR discrim* OR diagnos*) AND (photograph* OR imag*) AND (“auto* detect*” OR detect* OR predict*) AND (segment*))

7.1.4. *Database: MEDLINE*

(glaucoma and (fundus or retinal) and (classif* or discrim* or diagnos*) and (photograph* or imag*) and (auto* detect* or detect* or predict*) and segment*).tw.

7.2. *Eligibility criteria*

We included papers if:

1. The paper uses segmented fundus images of the Optic Nerve Head (ONH).
2. The paper proposes a methodology/framework for the classification of glaucoma.
3. Full text is available online.
4. Full text is available in English.

We excluded papers if:

1. Interested purely in segmentation of fundus images (provide no classification of glaucoma following segmentation).
2. Focused purely on classification via methods that require no segmentation of the fundus image (i.e., one step AI frameworks).

8. Disclosure

This work was supported by a PhD studentship funding from the British Council for Prevention of Blindness to Dr. Gabriela Czanner, Dr. Silvester Czanner, Professor Colin E. Willoughby, and Dr. Srinivasan Kavitha. The authors have no conflicts of interest to disclose.

Permission requirements

Not applicable.

REFERENCES

1. Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. Information Fusion. 2021;76:243–97.
2. Abdullah F, Imtiaz R, Madni HA, Khan HA, Khan TM, Khan MAU, et al. A review on glaucoma disease detection using computerized techniques. IEEE Access. 2021;9:37311–33.
3. Afolabi OJ, Mabuza-Hocquet GP, Nelwamondo FV, Paul BS. The Use of U-Net Lite and extreme gradient boost (XGB) for glaucoma detection. IEEE Access. 2021;9:47411–24.
4. Agarwal A, Gulia S, Chaudhary S, Dutta MK, Burget R, Riha K. Automatic glaucoma detection using adaptive threshold based technique in fundus image. In: 2015 38th International Conference on Telecommunications and Signal Processing (TSP); 2015. p. 416–20.
5. Agarwal A, Gulia S, Chaudhary S, Dutta MK, Travieso CM, Alonso-Hernández JB. A novel approach to detect glaucoma in retinal fundus images using cup-disk and rim-disk ratio. In: 2015 4th International Work Conference on Bioinspired Intelligence (IWOBI); 2015. p. 139–44.
6. Ahmad J, Muhammad J, Aziz L, Ayub S, Akram M, Basit I. Glaucoma detection through optic disc and cup segmentation using K-mean clustering. In 2016. p. 143–7.
7. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ. 2019;7:e7702.
8. Akram MU, Tariq A, Khalid S, Javed MY, Abbas S, Yasin UU. Glaucoma detection using novel optic disc localization, hybrid feature set and classification techniques. Australas Phys Eng Sci Med. 2015;38(4):643–55.
9. Almazroa A, Burman R, Raahemifar K, Lakshminarayanan V. Optic disc and optic cup segmentation methodologies for glaucoma image detection: a survey. J Ophthalmol. 2015;2015:1–28.
10. Azuara-Blanco A, Banister K, Boachie C, McMeekin P, Gray J, Burr J, et al. Automated imaging technologies for the diagnosis of glaucoma: a comparative diagnostic study for the evaluation of the diagnostic accuracy, performance as triage tests and cost-effectiveness (GATE study). Health Technol Assess. 2016;8:1–168.
11. Barikian A, Haddock LJ. Smartphone assisted fundus fundoscopy/photography. Curr Ophthalmol Rep. 2018;6(1):46–52.
12. Batista FJF, Diaz-Aleman T, Sigut J, Alayon S, Arnay R, Angel-Pereira D. RIM-ONE DL: a unified retinal image database for assessing glaucoma using deep learning. Image Anal Stereol. 2020;39(3):161–7.
13. Bock R, Meier J, Nyúl LG, Hornegger J, Michelson G. Glaucoma risk index: automated glaucoma detection from color fundus images. Med Image Anal. 2010;14(3):471–81.
14. Božić-Štulić D, Braovic M, Stipanicev D. Deep learning based approach for optic disc and optic cup semantic segmentation for glaucoma analysis in retinal fundus images. 2020.
15. Budai A, Bock R, Maier A, Hornegger J, Michelson G. Robust vessel segmentation in fundus images. Int J Biomed Imaging. 2013;2013:e154860.
16. Burr JM, Mowatt G, Hernández R, Siddiqui M aR, Cook J, Lourenco T, et al. The clinical effectiveness and cost-effectiveness of screening for open angle glaucoma: a systematic review and economic evaluation. Health Technol Assess Winch Engl. 2007;11(41) iii–iv, ix–x, 1–190.

17. Camara J, Neto A, Pires IM, Villasana MV, Zdravevski E, Cunha A. Literature review on artificial intelligence methods for glaucoma screening, segmentation, and classification. J Imaging. 2022;8(2):19.
18. Carmona EJ, Rincón M, García-Feijoó J, Martínez-de-la-Casa JM. Identification of the optic nerve head with genetic algorithms. Artif Intell Med. 2008;43(3):243–59.
19. Chefer H, Gur S, Wolf L. Transformer Interpretability Beyond Attention Visualization. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) [Internet]. Nashville, TN, USA: IEEE; 2021. p. 782–91. [cited 2022 Jun 20]Available from:.
20. Chen C, Li O, Tao D, Barnett A, Rudin C, Su JK. This Looks Like That: Deep Learning for Interpretable Image Recognition. Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2019. [cited 2022 Jun 20]. Available from: https://proceedings.neurips.cc/paper/2019/hash/adf7ee2dcf142b0e11888e72b43fcb75-Abstract.html .
21. Cheng KKW, Tatham AJ. Spotlight on the disc-damage likelihood scale (DDLS). Clin Ophthalmol Auckl NZ. 2021;15:4059–71.
22. Civit-Masot J, Domínguez-Morales MJ, Vicente-Díaz S, Civit A. Dual machine-learning system to aid glaucoma diagnosis using disc and cup feature extraction. IEEE Access. 2020;8:127519–29.
23. Czanner G, Bunce C. Chapter 10 - Statistical analysis and design in ophthalmology: Toward optimizing your data.Trucco E, MacGillivray T, Xu Y (eds.). Computational Retinal Image Analysis. Academic Press; 2019. p. 171–97.
24. Das P, Nirmala S, Medhi J. Detection of glaucoma using Neuroretinal Rim information. In 2016. p. 181–6.
25. Das P, Nirmala SR, Medhi JP. Diagnosis of glaucoma using CDR and NRR area in retina images. Netw Model Anal Health Inform Bioinforma. 2016;5(1):3.
26. Das S, Kuht HJ, De Silva I, Deol SS, Osman L, Burns J, et al. Correction: feasibility and clinical utility of handheld fundus cameras for retinal imaging. Eye. 2022:1–2. https://www.nature.com/articles/s41433-021-01926-y#citeas.
27. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med. 2018;24(9):1342–50.
28. Decencière E, Zhang X, Cazuguel G, Lay B, Cochener B, Trone C, et al. Feedback on a publicly distributed image database: the messidor database. Image Anal Stereol. 2014;33(3):231.
29. Deepika E, Maheswari S. Earlier glaucoma detection using blood vessel segmentation and classification. In: 2018 2nd International Conference on Inventive Systems and Control (ICISC). 2018. p. 484–90.
30. Delgado MF, Abdelrahman AM, Terahi M, Miro Quesada Woll JJ, Gil-Carrasco F, Cook C, et al. Management of glaucoma in developing countries: challenges and opportunities for improvement. Clin Outcomes Res CEOR. 2019;11:591–604.
31. Dutta K, Mukherjee R, Kundu S, Biswas T, Sen A. Automatic Evaluation and Predictive Analysis of Optic Nerve Head for the Detection of Glaucoma. In: 2018 2nd International Conference on Electronics, Materials Engineering Nano-Technology (IEMENTech). 2018. p. 1–7.
32. Dutta MK, Mourya AK, Singh A, Parthasarathi M, Burget R, Riha K. Glaucoma detection by segmenting the super pixels from fundus colour retinal images. In: 2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom). 2014. p. 86–90.
33. Faes L, Liu X, Wagner SK, Fu DJ, Balaskas K, Sim DA, et al. A clinician's guide to artificial intelligence: how to critically appraise machine learning studies. Transl Vis Sci Technol. 2020;9(2):1–7.
34. Foong AWP, Saw S-M, Loo J-L, Shen S, Loon S-C, Rosman M, et al. Rationale and methodology for a population-based study of eye diseases in Malay people: the singapore malay eye study (SiMES). Ophthalmic Epidemiol. 2007;14(1):25–35.
35. Fumero F, Alayón S, Sanchez JL, Sigut J, Gonzalez-Hernandez M. RIM-ONE: An open retinal image database for optic nerve evaluation. In 2011. p. 1–6.
36. Fumero F, Sigut J, Alayón S, González-Hernández M, González M. Interactive Tool and Database for Optic Disc and Cup Segmentation of Stereo and Monocular Retinal Fundus Images. 2015;7.
37. Garway-Heath DF, Crabb DP, Bunce C, Lascaratos G, Amalfitano F, Anand N, et al. Latanoprost for open-angle glaucoma (UKGTS): a randomised, multicentre, placebo-controlled trial. Lancet Lond Engl. 2015;385(9975):1295–304.
38. George D, Lehrach W, Kansky K, Lázaro-Gredilla M, Laan C, Marthi B, et al. A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs. Science. 2017;358(6368).
39. Giancardo L, Meriaudeau F, Karnowski TP, Li Y, Garg S, Tobin KW, et al. Exudate-based diabetic macular edema detection in fundus images using publicly available datasets. Med Image Anal. 2012;16(1):216–26.
40. Gordon MO, Beiser JA, Brandt JD, Heuer DK, Higginbotham EJ, Johnson CA, et al. The ocular hypertension treatment study: baseline factors that predict the onset of primary open-angle glaucoma. Arch Ophthalmol Chic Ill 1960. 2002;120(6):714–20 discussion 829-830.
41. Gunn P. Glaucoma management part 2 - Optic disc assessment in glaucoma. Opt Sel. 2016;2016(3):118–21.
42. Hamid S, Desai P, Hysi P, Burr JM, Khawaja AP. Population screening for glaucoma in UK: current recommendations and future directions. Eye. 2021;36:1–6. doi:10.1038/s41433-021-01687-8.
43. Harper R, Reeves B, Smith G. Observer variability in optic disc assessment: implications for glaucoma shared care. Ophthalmic Physiol Opt. 2014;20(4):265–73.
44. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York, NY: Springer New York; 2009.
45. Hemelings R, Elen B, Barbosa-Breda J, Blaschko MB, De Boever P, Stalmans I. Deep learning on fundus images detects glaucoma beyond the optic disc. Sci Rep. 2021;11(1):20313.
46. Issac A, Dutta M. Automated framework for screening of glaucoma through cloud computing. J Med Syst. 2019;43:136.
47. Issac A, Partha Sarathi M, Dutta MK. An adaptive threshold based image processing technique for improved glaucoma detection and classification. Comput Methods Programs Biomed. 2015;122(2):229–44.
48. Jones PR, Campbell P, Callaghan T, Jones L, Asfaw DS, Edgar DF, et al. Glaucoma home monitoring using a tablet-based visual field test (Eyecatcher): an assessment of accuracy and adherence over 6 months. Am J Ophthalmol. 2021;223:42–52.
49. Kang H, Li X, Su X. Cup-disc and retinal nerve fiber layer features fusion for diagnosis glaucoma. 2020 Mar 1;11314:113143Z.
50. Karkuzhali S, Manimegalai D. Computational intelligence-based decision support system for glaucoma detection. Biomed Res. 2017;28(11):12.
51. Kausu TR, Gopi V, Wahid K, Doma W, Niwas S I. Combination of clinical and multiresolution features for glaucoma detection and its classification using fundus images. Biocybern Biomed Eng. 2018;38(2):38.

52. Khalil T, Akram MU, Khalid S, Jameel A. Improved automated detection of glaucoma from fundus image using hybrid structural and textural features. IET Image Proc. 2017;11(9):693–700.
53. Khan SM, Liu X, Nath S, Korot E, Faes L, Wagner SK, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. Lancet Digit Health. 2021;3(1):e51–66.
54. Krishna Adithya V, Williams BM, Czanner S, Kavitha S, Friedman DS, Willoughby CE, et al. EffUnet-SpaGen: an efficient and spatial generative approach to glaucoma detection. J Imag. 2021;7(6):92.
55. Krishnan R, Sekhar V, Sidharth J, Gautham S, Gopakumar G. Glaucoma Detection from Retinal Fundus Images. In: 2020 International Conference on Communication and Signal Processing (ICCSP); 2020. p. 0628–31.
56. Lam PY, Chow SC, Lai JSM, Choy BNK. A review on the use of telemedicine in glaucoma and possible roles in COVID-19 outbreak. Surv Ophthalmol. 2021;66(6):999–1008.
57. Li Z, He Y, Keel S, Meng W, Chang RT, He M. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. Ophthalmology. 2018;125(8):1199–206.
58. Liu X, Song L, Liu S, Zhang Y. A review of deep-learning-based medical image segmentation methods. Sustainability. 2021;13(3):1224.
59. Lotankar M, Noronha K, Koti J. Detection of optic disc and cup from color retinal images for automated diagnosis of glaucoma. In: 2015 IEEE UP Section Conference on Electrical Computer and Electronics (UPCON); 2015. p. 1–6.
60. Soorya M, Issac A, Dutta MK. An automated and robust image processing algorithm for glaucoma diagnosis from fundus images using novel blood vessel tracking and bend point detection. Int J Med Inf. 2018;110:52–70.
61. MacCormick IJC, Williams BM, Zheng Y, Li K, Al-Bander B, Czanner S, et al. Accurate, fast, data efficient and interpretable glaucoma diagnosis with automated spatial analysis of the whole cup to disc profile. Bhattacharya S, editor. PLOS ONE. 2019;14(1):e0209409.
62. Mansour RF, Al-Marghilnai A. Glaucoma detection using novel perceptron based convolutional multi-layer neural network classification. Multidim Syst Sign Proc. 2021;32(4):1217–35.
63. McCann P, Hogg R, Wright DM, Pose-Bazarra S, Chakravarthy U, Peto T, et al. Glaucoma in the northern Ireland cohort for the longitudinal study of ageing (NICOLA): cohort profile, prevalence, awareness and associations. Br J Ophthalmol. 2020;104(11):1492–9.
64. Mirzania D, Thompson AC, Muir KW. Applications of deep learning in detection of glaucoma: a systematic review. Eur J Ophthalmol. 2021;31(4):1618–42.
65. Mukherjee R, Kundu S, Dutta K, Sen A, Majumdar S. predictive diagnosis of glaucoma based on analysis of focal notching along the neuro-retinal rim using machine learning. Pattern Recognit Image Anal. 2019;29(3):523–32.
66. Mursch-Edlmayr AS, Ng WS, Diniz-Filho A, Sousa DC, Arnould L, Schlenker MB, et al. Artificial intelligence algorithms to diagnose glaucoma and detect glaucoma progression: translation to clinical practice. Transl Vis Sci Technol. 2020;9(2):55. doi:10.1167/tvst.9.2.55.
67. Mvoulana A, Kachouri R, Akil M. Fully automated method for glaucoma screening using robust optic nerve head detection and unsupervised segmentation based cup-to-disc ratio computation in retinal fundus images. Comput Med Imaging Graph. 2019;77:77.
68. Myers JS, Fudemberg SJ, Lee D. Evolution of optic nerve photography for glaucoma screening: a review. Clin Experiment Ophthalmol. 2018;46(2):169–76.
69. Narasimhan K, Vijayarekha DK. An efficient automated system for glaucoma detection using fundus image. 2011;33:7.
70. Neto A, Camera J, Oliveira S, Cláudia A, Cunha A. Optic disc and cup segmentations for glaucoma assessment using cup-to-disc ratio. Proc Comput Sci. 2022;196:485–92.
71. Odstrcilik J, Kolar R, Budai A, Hornegger J, Jan J, Gazarek J, et al. Retinal vessel segmentation by improved matched filtering: evaluation on a new high-resolution fundus image database. IET Image Proc. 2013;7(4):373–83.
72. Ong EP, Cheng J, Wong DWK, Tay ELT, Teo HY, Grace Loo R, et al. Automatic Glaucoma Detection from Stereo Fundus Images. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC); 2020. p. 1540–3.
73. Pathan S, Kumar P, Pai RM, Bhandary SV. Automated segmentation and classifcation of retinal features for glaucoma diagnosis. Biomed Sign Proc Control. 2021;63:102244.
74. Perdomo O, Andrearczyk V, Meriaudeau F, Müller H, González FA, et al. Glaucoma diagnosis from eye fundus images based on deep morphometric feature estimation.Stoyanov D, Taylor Z, Ciompi F, Xu Y, Martel A, Maier-Hein L, et al. (eds.). Computational Pathology and Ophthalmic Medical Image Analysis. Cham: Springer International Publishing; 2018. p. 319–27.
75. Poon LY-C, Solá-Del Valle D, Turalba AV, Falkenstein IA, Horsley M, Kim JH, et al. The ISNT Rule: how often does it apply to disc photographs and retinal nerve fiber layer measurements in the normal population? Am J Ophthalmol. 2017;184:19–27.
76. Quigley HA, Broman AT. The number of people with glaucoma worldwide in 2010 and 2020. Br J Ophthalmol. 2006;90(3):262–7.
77. Raja PMS, Ramanan K. Damped least-squares recurrent deep neural learning classification for glaucoma detection. In: 2019 International Conference on Data Science and Engineering (ICDSE); 2019. p. 160–5.
78. Rodriguez-Una I, Azuara-Blanco A. New technologies for glaucoma detection. Asia-Pac J Ophthalmol. 2018;7(6):394–404.
79. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15.
80. Sarhan MH, Nasseri MA, Zapp D, Maier M, Lohmann CP, Navab N, et al. Machine learning techniques for ophthalmic data processing: a review. IEEE J Biomed Health Inform. 2020;24(12):3338–50.
81. Saxena R, Singh D, Vashist P. Glaucoma: an emerging peril. Indian J Community Med. 2013;38(3):135.
82. Singh LK, Pooja Garg H, Khanna M, Bhadoria RS. An enhanced deep image model for glaucoma diagnosis using feature-based detection in retinal fundus. Med Biol Eng Comput. 2021;59(2):333–53.
83. Sivaswamy J, Krishnadas S, Chakravarty A, Gopal D, Joshi G, Ujjwal, et al. JSM Biomedical Imaging Data Papers. A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. JSM Biomed Imaging Data Pap. 2015;2(1):1004.
84. Sivaswamy J, Krishnadas SR, G Datt Joshi, Jain M, Drishti-GS Syed Tabish AU. Retinal image dataset for optic nerve head(ONH) segmentation. In: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI); 2014. p. 53–6.
85. Spaeth GL, Henderer J, Liu C, Kesen M, Altangerel U, Bayer A, et al. The disc damage likelihood scale: reproducibility of a new method of estimating the amount of optic nerve damage caused by glaucoma. Trans Am Ophthalmol Soc. 2002;100:181–6.

86. Steinmetz JD, Bourne RRA, Briant PS, Flaxman SR, Taylor HRB, Jonas JB, et al. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. Lancet Glob Health. 2021;9(2):e144–60.
87. Stella Mary MCV, Rajsingh EB, Naik GR. Retinal fundus image analysis for diagnosis of glaucoma: a comprehensive survey. IEEE Access. 2016;4:4327–54.
88. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. Circ Cardiovasc Qual Outcomes. 2020;13(10):e006556.
89. Thakur N, Juneja M. Survey on segmentation and classification approaches of optic cup and optic disc for diagnosis of glaucoma. Biomed Sign Proc Control. 2018;42:162–89.
90. Tham Y-C, Li X, Wong TY, Quigley HA, Aung T, Cheng C-Y. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. Ophthalmology. 2014;121(11):2081–90.
91. Thompson AC, Jammal AA, Medeiros FA. A Review of Deep Learning for Screening, Diagnosis, and Detection of Glaucoma Progression. Transl Vis Sci Technol. 2020;9(2):42.
92. Upadhyaya S, Agarwal A, Rengaraj V, Srinivasan K, Newman Casey PA, Schehlein E. Validation of a portable, non-mydriatic fundus camera compared to gold standard dilated fundus examination using slit lamp biomicroscopy for assessing the optic disc for glaucoma. Eye. 2021;36:1–7.
93. Van Calster B, McLernon D, van Smeden M, Wynants L, Steyerberg E. Calibration: The Achilles heel of predictive analytics. BMC Med. 2019;36:17.
94. Veena HN, Muruganandham A, Kumaran TS. A Review on the optic disc and optic cup segmentation and classification approaches over retinal fundus images for detection of glaucoma. SN Appl Sci. 2020;2(9):1476.
95. Vijapur NA, Kunte RSR. Sensitized glaucoma detection using a unique template based correlation filter and undecimated isotropic wavelet transform. J Med Biol Eng. 2017;37(3):365–73.
96. Xu Y, Hu M, Liu H, Yang H, Wang H, Lu S, et al. A hierarchical deep learning approach with transparency and interpretability based on small samples for glaucoma diagnosis. npj Digit Med. 2021;4(1):1–11.
97. Yunitasari DA, Sigit R, Harsono T. Glaucoma detection based on cup-to-disc ratio in retinal fundus image using support vector machine. In: 2021 International Electronics Symposium (IES); 2021. p. 368–73.
98. Zahoor MN, Fraz MM. A correction to the article "fast optic disc segmentation in retina using polar transform. IEEE Access. 2018;6:4845–9.
99. Zhang Z, Yin F, Liu J, Wong W, Tan N, Lee B-H, et al. ORIGA(-light): An Online Retinal Fundus Image Database for Glaucoma Analysis and Research. In 2010. P. 3065–8.
100. Zulfira FZ, Suyanto S, Septiarini A. Segmentation technique and dynamic ensemble selection to enhance glaucoma severity detection. Computers in Biol Med. 2021;139:104951.

OTHER REFERENCES

A. Vision impairment and blindness [Internet]. [cited 2021 Nov 14]. Available from: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment
B. Hospital Outpatient Activity - 2014-15 [Internet]. NHS Digital. [cited 2021 Dec 14]. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/hospital-outpatient-activity/hospital-outpatient-activity-2014-15
C. Leslie D . Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector [Internet]. Zenodo; 2019 [cited 2021 Dec 17]. Available from: https://zenodo.org/record/3240529
D. U.S. Food & Drug Administration. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. Case Medical Research [Internet]. 2018 [cited 2022 Jun 20]; Available from: https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye
E. McNeil R. Coming to terms with AI [Internet]. Eye News. [cited 2021 Aug 6]. Available from: https://www.eyenews.uk.com/features/ophthalmology/post/coming-to-terms-with-ai
F. Spiegelhalter D. Should We Trust Algorithms? Harv Data Sci Rev [Internet]. 2020 [cited 2021 Dec 17];2(1). Available from: https://hdsr.mitpress.mit.edu/pub/56lnenzj
G. Art. 15 GDPR – Right of access by the data subject - General Data Protection Regulation (GDPR) [Internet]. General Data Protection Regulation (GDPR). [cited 2022 Jun 20]. Available from: https://gdpr-info.eu/art-15-gdpr/
H. Centre for Information Policy Leadership (CIPL). Artificial Intelligence and Data Protection How the GDPR Regulates AI [Internet]. Informationpolicycentre.com. 2020 [cited 2022 Jun 20]. Available from: https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl-hunton_andrews_kurth_legal_note_-_how_gdpr_regulates_ai__12_march_2020_.pdf
I. High-level expert group on artificial intelligence | Shaping Europe's digital future [Internet]. [cited 2022 Jun 20]. Available from: https: //digital-strategy.ec.europa.eu/en/policies/expert-group-ai
J. Explainable AI: the basics POLICY BRIEFING [Internet]. Royalsociety.org. 2019 [cited 2022 Jun 20]. Available from: https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf
K. Reporting guidelines | The EQUATOR Network [Internet]. [cited 2021 Nov 14]. Available from: https://www.equator-network.org/reporting-guidelines/
L. . European Glaucoma Society. Terminology and Guidelines for Glaucoma [Internet]. 5th ed. European Glaucoma Society; 2020. [cited 2021 Jul 26]. Available from: https://www.eugs.org/eng/guidelines.asp .
M. Kauppi T. DIARETDB1 - standard diabetic retinopathy database [Internet]. DIARETDB1 - Standard Diabetic Retinopathy Database Calibration level 1. 2007 [cited 2021 Jul 15]. Available from: https://www.it.lut.fi/project/imageret/diaretdb1/