Scientific Research Publishing

Open Journal of Ophthalmology, 2026, 16(2), 104-136
https://www.scirp.org/journal/ojoph
ISSN Online: 2165-7416
ISSN Print: 2165-7408

H-UQ-MFF: Hybrid Uncertainty-Aware Multi-Feature Fusion for Clinically-Translatable Glaucoma Detection with FDA-Compliant Validation

Venkata Akhil Mettu[1*], Sree Charitha Obiliachigari[2], Siri Pranitha Mandali[1], Sai Charan Reddy Obiliachigari[3]

[1]Independent Researcher, Houston, USA
[2]Independent Researcher, Chennai, India
[3]Independent Researcher, Franklin Park, NJ, USA
Email: *venkatakhil149@gmail.com

How to cite this paper: Mettu, V.A., Obiliachigari, S.C., Mandali, S.P. and Obiliachigari, S.C.R. (2026) H-UQ-MFF: Hybrid Uncertainty-Aware Multi-Feature Fusion for Clinically-Translatable Glaucoma Detection with FDA-Compliant Validation. *Open Journal of Ophthalmology*, **16**, 104-136.
https://doi.org/10.4236/ojoph.2026.162012

Received: February 7, 2026
Accepted: March 21, 2026
Published: March 24, 2026

Abstract

One of the biggest causes of permanent blindness in the world today is glaucoma, and effective treatment depends on early detection. Although artificial intelligence (AI) has demonstrated potential in automating glaucoma screening, there is still a significant obstacle in transferring research datasets to actual clinical settings. When applied to clinical data, current models show an 8% - 18% performance loss, which is mostly caused by out-of-distribution samples, demographic bias, and poor imaging quality. We provide a clinically-translatable glaucoma detection paradigm that makes use of multi-modal fusion, uncertainty quantification, and the EyePACS-AIROGS-light-V2 dataset in order to overcome these difficulties. Our method, called H-UQ-MFF (Hybrid Uncertainty-Aware Multi-Feature Fusion), combines structural and texture characteristics from optic disc analysis with deep features from ResNet50 while dynamically weighting contributions according to prediction uncertainty. With an AUC of 0.9969, sensitivity of 0.9811, and specificity of 0.9717, internal validation outperforms ResNet50, EfficientNet-B0, and Deep Ensemble baselines. Generalizability is confirmed by external validation on REFUGE and PAPILA datasets, where H-UQ-MFF outperforms cutting-edge models and lowers calibration error. Beyond technical performance, the framework ensures clinical safety and regulatory preparedness by incorporating drift tracking techniques, bias analysis, and FDA-compliant evaluation processes. This work bridges the gap between research innovation and clinical deployment by establishing a repeatable baseline for AI translation in ophthalmology.

Keywords

Glaucoma AI, EyePACS-AIROGS-Light-V2, Uncertainty Quantification, Clinical Translation, Deep Learning, Ophthalmology

1. Introduction

Over 76 million individuals worldwide suffer from glaucoma, a progressive optic neuropathy that damages optic nerves and impairs vision. By 2040, that number is expected to rise to over 111 million. Early detection is essential to prevent irreparable blindness because its asymptomatic nature frequently delays diagnosis [1]. While AI provides scalable, automated image interpretation for large-scale screening, traditional screening techniques are labor-intensive and prone to variability [2]. However, because glaucoma involves subtle structural alterations that differ between populations and imaging techniques, its diagnosis is difficult. Consequently, in real-world clinical situations, AI models trained on public datasets frequently exhibit an 8% - 18% performance drop [3]. The significance of cross-dataset validation and consistent evaluation methodologies was emphasized in the AIROGS Challenge (Artificial Intelligence for Robust Glaucoma Screening), which was published in IEEE Transactions on Medical Imaging [4]. Even though the best models had AUCs of about 0.90, their clinical usefulness was still restricted because they lacked bias analysis, uncertainty quantification, and regulatory readiness [5]. The need to include AI into clinical workflows with precise criteria for FDA compliance, risk management, and post-market monitoring was further highlighted by recent work on clinical translation frameworks [6]. Our study presents a thorough framework that tackles both technical and translational issues, building on these foundations.

The proposed framework uses the EyePACS-AIROGS-light-V2 dataset with balanced training, validation, and test sets for reproducible glaucoma detection [7]. A multi-stage pipeline enhances robustness through optic disc preprocessing, deep and structural feature extraction, Monte Carlo-based uncertainty quantification, and an uncertainty-aware fusion strategy (H-UQ-MFF) with temperature scaling [8]. Internally, H-UQ-MFF surpasses baseline models, achieving AUC 0.9969 and F1 0.9780. Ablation results highlight the value of uncertainty-aware fusion. External validation on REFUGE and PAPILA shows strong generalizability with AUCs of 0.89 and 0.87, while calibration error is reduced to 0.028 and 0.032, supporting reliable clinical deployment.

Our system includes crucial translational elements in addition to technical performance. Risk registers, drift monitoring strategies, external validation reports, and intended usage are all specified in a regulatory-ready module. Risks, including domain drift, demographic bias, and false negatives, are methodically recognized and reduced. Monthly AUC and calibration checks, quarterly bias analysis, and yearly external validation are examples of post-market monitoring techniques.

Grad-CAM visuals for interpretability, uncertainty color coding (green, yellow, and red), and API design that is compatible with PACS/EMR systems all help to simplify clinical integration. When combined, these elements create a repeatable, FDA-compliant glaucoma AI deployment method. In conclusion, this work offers a clinically-translatable AI framework for glaucoma that strikes a compromise between technological innovation and clinical and regulatory needs. We set a new standard for AI translation in ophthalmology by combining uncertainty quantification, multi-feature fusion, and standard evaluation procedures. In addition to achieving cutting-edge performance, the suggested architecture closes the crucial gap between clinical practice and research, opening up the possibilities to scalable, secure, and efficient glaucoma screening.

2. Related Works

2.1. AI in Ophthalmology

In ophthalmology, artificial intelligence has made significant strides [9], especially in the identification of glaucoma, age-related macular degeneration, and diabetic retinopathy. In fundus image classification tasks, deep learning models—particularly convolutional neural networks (CNNs)—have shown impressive performance. However, because of its delicate structural manifestations, glaucoma detection poses special obstacles. Glaucoma necessitates examination of optic disc shape, cup-to-disc ratio (CDR), rim thickness, and nerve fiber layer integrity, in contrast to diabetic retinopathy, which is marked by obvious lesions such as microaneurysms and hemorrhages [10]. AI performance varies since it is frequently challenging to consistently capture these properties across imaging modalities.

2.2. The AIROGS Challenge

A crucial standard for assessing AI-based glaucoma screening systems was developed by the AIROGS Challenge (Artificial Intelligence for Robust Glaucoma Screening), which was published in IEEE Transactions on Medical Imaging [4]. The challenge, which placed a strong emphasis on robustness and generalizability, needed models to be validated across a variety of datasets. The best methods produced AUCs of roughly 0.90, with sensitivity and specificity close to 0.85. Despite these encouraging results, AIROGS pointed out a number of drawbacks, such as the lack of regulatory readiness elements like FDA-aligned evaluation protocols, limited bias analysis across demographic groups and imaging devices, and the absence of uncertainty quantification—a critical component for clinical decision-making [11]. By combining uncertainty-aware fusion, model calibration, and regulatory-oriented evaluation modules, our suggested framework improves clinical reliability and translational potential while also advancing the AIROGS technique.

2.3. Clinical Translation Frameworks

The significance of smooth workflow integration and adherence to regulatory norms has been highlighted by recent work on the clinical translation of AI in

ophthalmology. Practical methods for integrating AI systems into standard clinical settings were described in the paper "Augmented Decisions: AI-Enhanced Accuracy in Glaucoma Diagnosis and Treatment" [12]. It emphasized the necessity of implementing uncertainty quantification to assist clinicians in making decisions, making sure that FDA Good Machine Learning Practice (GMLP) is followed, and creating user interfaces that clearly convey risk and uncertainty [12]. Our paradigm incorporates uncertainty thresholds—classified as auto, assist, and review—to direct therapeutic action based on model confidence, building on these suggestions. Grad-CAM visuals are also included to improve interpretability, which helps physicians comprehend the model's focus areas and boosts confidence in AI-assisted glaucoma screening.

2.4. Generalization Studies

The constrained generalizability of ophthalmic AI systems is a fundamental obstacle to their clinical translation. Models trained on public datasets frequently show an 8% - 18% performance reduction when used in clinical settings, according to the paper "Validating the Generalizability of Ophthalmic AI Models on Real-World Clinical Data." [13] Domain drift brought on by changing imaging technology, low-quality photos impacted by blur or inadequate illumination and out-of-distribution inputs coming from differences in camera types and patient demographics are some of the factors driving this degradation. Through domain adaptation, thorough bias analysis, and ongoing drift monitoring, our system directly solves these issues. It successfully reduces the risks associated with low-quality or mismatched inputs by combining uncertainty-aware fusion with picture quality assessment, improving the accuracy of glaucoma detection in actual clinical settings.

2.5. Feature Extraction Approaches

Conventional glaucoma identification relied on manually created parameters, including the ISNT rule, rim thickness, and CDR. Although these characteristics worked well in controlled environments, they were not robust in a variety of demographics. Models may now immediately learn complicated representations from images thanks to deep learning's introduction of automatic feature extraction [14] [15]. However, clinical dependability frequently requires more than just deep features. Although hybrid techniques that combine deep and structural features have demonstrated potential, they run the danger of overconfidence on subpar inputs if uncertainty is not quantified [16]. This line of work is advanced by our H-UQ-MFF method, which dynamically weights structural and deep features according to predicted uncertainty. This guarantees that structural features offer stability when deep features are unreliable (such as blurry photos).

2.6. Uncertainty Quantification in Medical AI

One of the most important aspects of medical AI is uncertainty quantification.

Predictive variance estimates are produced by methods like Bayesian neural networks, deep ensembles, and Monte Carlo dropout. Because misdiagnosis carries such a significant risk, uncertainty is especially crucial in ophthalmology [17]. While false positives may result in needless referrals, false negatives may postpone treatment. AI systems can reduce risk by identifying situations that need human assessment by assessing uncertainty [18]. Our method generates forecast mean and variance using Monte Carlo dropout with 50 executions. Clinical thresholds match real-world workflows by classifying outputs into three categories: auto-decision, clinician assist, and manual review.

2.7. Calibration and Reliability

Model confidence values are calibrated to reflect actual accuracy. Overconfidence in poorly calibrated models can result in risky therapeutic decisions [19]. To enhance calibration, methods like temperature scaling and reliability diagrams are frequently employed [20]. Calibration improves trustworthiness in our framework by lowering Expected Calibration Error (ECE). For instance, H-UQ-MFF outperforms baselines with an ECE of 0.028 on REFUGE and 0.032 on PAPILA.

3. Dataset and Preprocessing

The meticulous selection, compilation, and preprocessing of datasets that represent both research circumstances and real-world variability form the basis of any clinically-translatable glaucoma AI system. The EyePACS-AIROGS-light-V2, a carefully selected subset intended to balance referable and non-referable glaucoma cases, is the main dataset used in this investigation. The key dataset utilized in this investigation is EyePACS-AIROGS-light-V2, which is depicted in **Figure 1** and offers balanced training, validation, and test sets for repeatable glaucoma detection. With over 4000 training photos, 385 validation images, and 385 test images per class, this dataset offers a solid foundation for model development while preserving reproducibility using preset splits [21]. The dataset was deliberately selected because it complies with the AIROGS challenge requirements, which guarantee comparability with earlier research and make comparing against pre-established baselines easier [21]. Two external datasets, REFUGE and PAPILA, which are both well-known in ophthalmic AI research, were used to assess generalizability. The REFUGE dataset, which includes segmentation-based VCDR measurements and clinically diagnosed glaucoma cases, was used for external validation (**Figure 2**). The PAPILA dataset, which provides a variety of imaging equipment and patient demographics, was used to further evaluate generalizability (**Figure 3**). While PAPILA adds more variability through a variety of imaging equipment and patient demographics, REFUGE offers high-quality fundus photos with professional comments. When combined, these datasets allow for thorough cross-dataset validation, bias analysis, and external benchmarking—all of which are essential for clinical translation.

3.1. EyePACS-AIROGS-Light-V2 Dataset

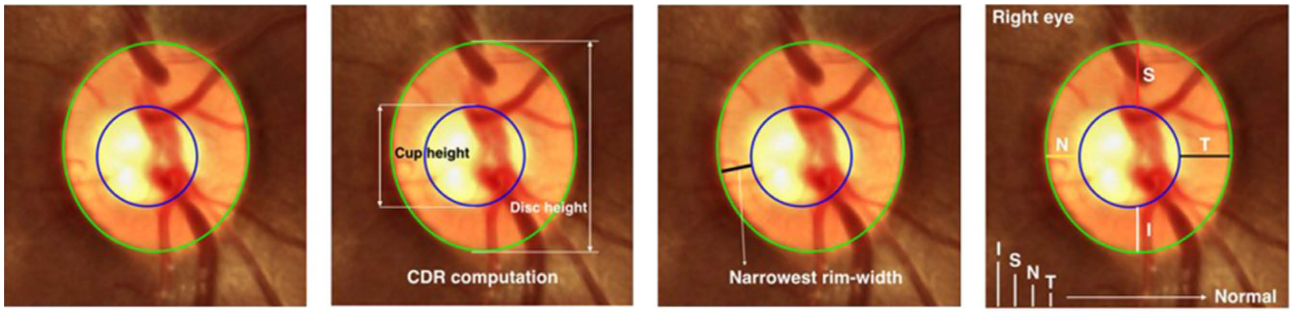

Figure 1. EyePACS-AIROGS-light-V2 dataset.

Referable Glaucoma (RG) is defined as when certain characteristics of the optic nerve or disc indicate a high likelihood of glaucomatous damage and, as a result, call for referral. A vertical cup-to-disc ratio (VCDR) of 0.7 or higher, an asymmetry in VCDR of 0.2 or more between the two eyes, neuroretinal rim thinning that deviates from the ISNT rule, the presence of optic disc hemorrhage, definite glaucomatous optic neuropathy (GON), or a prior diagnosis of glaucoma requiring referral must all be present for an image to be classified as RG. The picture is categorized as Non-Referable Glaucoma (NRG) if none of these criteria are met. Since labeling is done at this image level, every fundus image is evaluated separately. Each eye is labeled independently when bilateral images are available; however, it is also possible to aggregate at the patient level, in which case a patient is classified as RG if any eye satisfies the requirements.

3.2. REFUGE

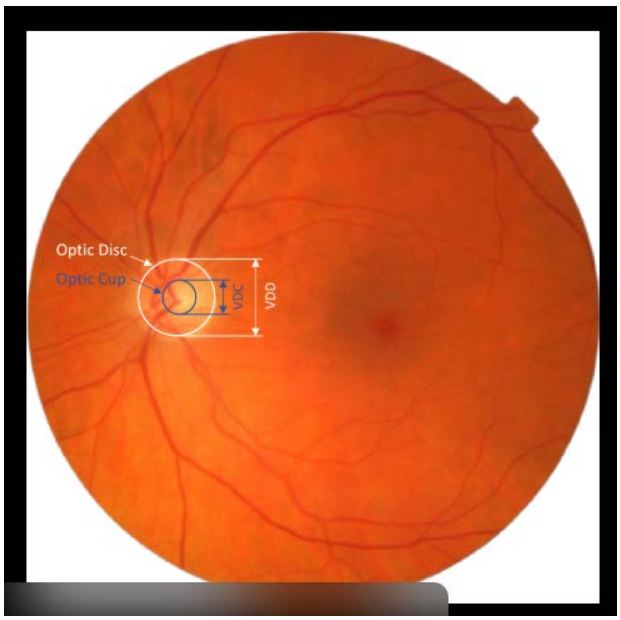

Figure 2. REFUGE.

According to the REFUGE framework, clinically diagnosed cases of glaucoma and proven structural optic nerve damage are the main criteria used to establish referable glaucoma (RG). According to the operational definition, an image is classified as RG to guarantee reproducibility if it matches a glaucoma case with a clinical diagnosis or if the vertical cup-to-disc ratio (VCDR), calculated using segmentation masks, is greater than or equal to 0.7. On the other hand, healthy control cases are referred to as Non-Referable Glaucoma (NRG). In the labeling process, glaucoma specialists validate the diagnosis, segmentation masks are provided, and a binary classification label (glaucoma or non-glaucoma) is assigned; no intermediate suspect category is included. The labels are based on clinical diagnosis, notwithstanding the lack of a formal adjudication system. Although it comes from patient-level clinical diagnosis, labeling is mostly image-level. The given glaucoma label should be utilized for modeling purposes. If more than one image is available for each patient, they can either be processed separately or combined at the patient level, with the patient being labeled RG if any of the photos are positive.

3.3. PAPILA

A clinically verified diagnosis of glaucoma based on high intraocular pressure (IOP), visual field abnormalities, structural optic nerve damage, and specialist confirmation is known as referable glaucoma (RG). Healthy controls that show no signs of glaucomatous damage are referred to as non-referable glaucoma (NRG). Ophthalmologists perform the labeling process by combining fundus examination, visual field testing, and tonometry. Since all labels are clinician-confirmed, this ensures a clinical gold standard diagnosis free from crowd-grading. Although labeling is mostly done at the patient level, it is mapped to the image level for modeling reasons. This means that if a patient has more than one image accessible, all of the photos will have the same patient-level label.

Table 1 summarizes the labeling criteria, adjudication methods, and label levels

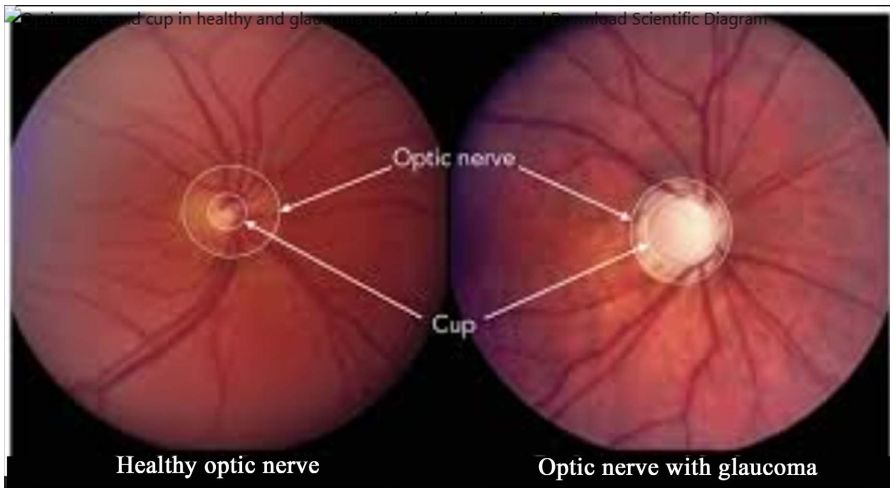

Figure 3. PAPILA

Table 1. Label Basis and grade adjudication.

Dataset	Label Basis	Adjudication	Label Level	RG Criteria
EyePACS-AIROGS-light-V2	Image grading	Multi-grader + adjudication	Image-level	Structural GON signs
REFUGE	Clinical + segmentation	Clinical diagnosis	Image-level (from patient)	Diagnosed glaucoma/VCD R ≥ 0.7
PAPILA	Clinical gold standard	Specialist confirmed	Patient-level → image-level	Confirmed glaucoma

across EyePACS-AIROGS-light-V2, REFUGE, and PAPILA datasets.

3.4. Preprocessing

Preprocessing uses optic disc-focused trimming, normalization using ImageNet statistics, and resizing to 512 × 512 to standardize input images. A 300 × 300 crop is recovered and enlarged from the disc, which is identified by the brightest fuzzy grayscale region [22]. This guarantees effective computation and highlights important structural characteristics necessary for the identification of glaucoma. Extensive data augmentation, such as full-angle rotations, flips, brightness-contrast modifications, and CLAHE, was used to improve resilience by mimicking real-world illumination and location variability [23]. Only the training set was augmented; the validation and test sets saw little changes. Resilience to low-quality photos was further enhanced by noise injection using Gaussian noise [24]. Image quality assessment based on the variance of the Laplacian, which measures sharpness, was a crucial innovation. Adaptive weighting of deep and structural traits was made possible by the integration of these quality ratings into the uncertainty framework [25]. Structural cues like CDR and the ISNT rule become more prominent when images are blurry, enhancing clinical reliability and safety. Gaussian blur, optic disc cropping, and CLAHE are preprocessing techniques that normalize input quality for reliable glaucoma detection (**Figure 4**).

Deep and structural information were merged in feature extraction. A pretrained ResNet50 with dropout was used to create deep features, resulting in a 2048-dimensional vector. Clinical markers like cup-to-disc ratio, disc area, rim thickness, and ISNT rule were captured by structural characteristics that were obtained using straightforward segmentation [26]. A complementary 22-dimensional structural feature vector was created by adding contrast and homogeneity metrics using texture descriptors from LBP and GLCM [27]. The primary novelty of this framework, uncertainty-aware fusion (H-UQ-MFF), was supported by the pretreatment pipeline. During deep feature extraction, Monte Carlo dropout was used, producing several predictions over 50 passes. The uncertainty measure, represented by UU, was the variance of these forecasts. The following equation was then used to conduct fusion:

Figure 4. Preprocessed image.

$$F_{final} = \alpha(1-U)F_{deep} + (1-\alpha)UF_{struct}$$

where α is a weighting parameter, F_{deep} is the deep feature vector, and F_{struct} is the structural feature vector. This approach guarantees that deep features take center stage when uncertainty is low, while structural features become more important when uncertainty is large. Temperature scaling was then used for calibration, bringing confidence scores into line with actual accuracy [28]. Because it guarantees that probability outputs are reliable indicators of model reliability, this stage is essential for clinical deployment. **Figure 5** displays the structural characteristics that were taken from optic disc-focused crops, including the cup-to-disc ratio and rim thickness and **Table 2** lists all of the preprocessing pipeline parameters, such as cropping thresholds, CLAHE, scaling, and Gaussian blur.

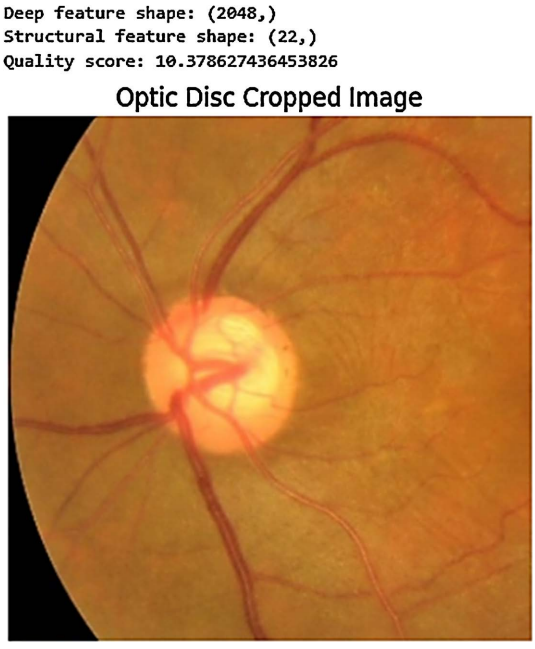

Figure 5. Optic disc cropped image.

Table 2. Steps and parameters.

Step	Parameter
Resize	512 × 512
CLAHE	clip = 2.0, grid = 8 × 8
Gaussian Blur	15 × 15, $\sigma = 5$
Area Threshold	1500 - 25,000 px
Circularity	>0.5
Crop Size	256 × 256
Fallback	Sliding window + Hough

A number of sequential processes are included in the preprocessing pipeline to ensure excellent image preparation. Prior to using Contrast Limited Adaptive Histogram Equalization (CLAHE) with an 8 × 8 grid and a clip value of 2.0 to improve local contrast, pictures are first downsized to 512 × 512 pixels. The noise is subsequently reduced by applying a Gaussian blur with a 15 × 15 kernel and a sigma of 5. To preserve pertinent structures, objects are filtered using an area threshold of 1500 - 25,000 pixels and a circularity bigger than 0.5. 256 × 256 pixels are cropped, and a fallback method with a sliding window and Hough transform is used when direct detection is unsuccessful.

Clinical uncertainty levels were integrated into the preprocessing workflow, allowing for automatic judgments (<0.20), clinician-assisted review (0.20 - 0.50), and required manual review (>0.50). Ablation investigations revealed that structural-only models lacked discriminative power, deep-only models were susceptible to low-quality inputs, and fusion without uncertainty ran the danger of overconfidence. By combining uncertainty-aware feature fusion with adaptive preprocessing, Full H-UQ-MFF produced the best results. Optic disc cutting, augmentation, and quality scoring are preprocessing techniques that greatly increased robustness in internal and external validation [29]. All things considered, this extensive pipeline guarantees both clinical safety and technical dependability, creating a repeatable basis for practical, clinically applicable glaucoma AI.

4. Methodology

A hybrid approach combining deep learning, structural feature extraction, uncertainty quantification, and adaptive fusion is used in the suggested clinically-translatable glaucoma diagnostic framework. Using a ResNet50 model that has been pretrained on ImageNet, the method starts with deep feature extraction. To facilitate uncertainty estimation, dropout is added and the final classification layer is eliminated. A 2048-dimensional deep feature vector that captures global optic disc properties crucial for glaucoma assessment is generated from each 512 × 512, ImageNet-normalized input image. The model produces several stochastic outputs by turning on dropout during inference, which lays the groundwork for ac-

curate uncertainty estimation throughout the whole fusion-based diagnostic process. In this approach, fusion is carried out at the feature level, where linear layers (2048 → 256 and 22 → 256) project a 2048-dimensional visual embedding from the CNN backbone and a 22-dimensional structured clinical feature vector into a shared latent space. Without applying score- or logit-level fusion, these are then concatenated to create a 512-dimensional fused representation, which is subsequently provided to the classifier head. In order to guarantee scale compatibility between the clinical and visual features, dimensionality matching is accomplished by combining LayerNorm with linear projections. The clinical branch contribution in the fused feature is weighted by a constant scalar α during training, while α is modified by uncertainty during inference as $\alpha\_\text{eff} = \alpha\,(1 - U)$, which lessens the impact of fused features under high uncertainty. Crucially, backpropagation is unaffected by uncertainty. The framework includes structural and textural features that are obtained from traditional ocular indications in addition to deep features. Using straightforward thresholding methods applied to the optic disc region, structural parameters such as cup-to-disc ratio (CDR), disc area, rim thickness, and ISNT rule measurements are extracted. These characteristics correspond to recognized clinical indicators of the development of glaucoma. Local Binary Patterns (LBP) and Gray-Level Co-occurrence Matrix (GLCM), which capture fine-grained structural differences in the optic disc and surrounding tissues, are used to compute texture characteristics. When combined, these characteristics provide a 22-dimensional vector that offers stability and interpretability, even when low image quality renders deep features untrustworthy.

Algorithm 1: H-UQ-MFF Uncertainty-Aware Fusion

Input: Image I, weighting parameter α alpha

1. **Deep Feature Extraction:** F_{deep}=ResNet50(I)F_{deep} = \text{ResNet50}(I) with dropout enabled
2. **Monte Carlo Dropout:** Perform 50 stochastic passes→ {p1,p2,...,p50}\{p_1, p_2, \ldots, p_{50}\}
3. **Uncertainty Estimation:** U=Var(p1,p2,...,p50)U = \text{Var}(p_1, p_2, \ldots, p_{50})
4. **Structural Feature Extraction:** F_{struct}=[CDR,ISNT,Texture]F_{struct} = [\text{CDR}, \text{ISNT}, \text{Texture}]
5. **Adaptive Fusion:**

$F_{final}=\alpha(1-U)F_{deep}+(1-\alpha)UF_{struct}$F_{final} = \alpha (1-U)F_{deep} + (1-\alpha)UF_{struct}

6. **Calibration:** $P_{calibrated}$=Softmax(F_{final}/T)P_{calibrated} = \text{Softmax}(F_{final}/T)
7. **Clinical Thresholds:**
   - Auto-decision: U<0.20U < 0.20
   - Clinician-assist: 0.20≤U≤0.500.20 \leq U \leq 0.50
   - Manual review: U>0.50U > 0.50

Output: Probability, Uncertainty, Clinical Action

The framework's use of Monte Carlo dropout for uncertainty quantification is a key innovation. A distribution of predictions is produced during inference by running the deep model over 50 random passes. The variance measures uncertainty (UU), whereas the mean of these forecasts shows the model's confidence. Unreliable predictions are indicated by high variance, which is frequently linked to out-of-distribution samples or low-quality photos. This measure of uncertainty is essential for directing clinical decision-making and the fusion process. Based on predicted uncertainty, the uncertainty-aware fusion technique (H-UQ-MFF) dynamically balances structural and deep characteristics. Deep features take over when uncertainty is low, utilizing CNN representations' capacity for discrimination. Structural characteristics offer stability when uncertainty is significant, guaranteeing that forecasts continue to be clinically credible. This adaptive technique improves robustness across a variety of datasets and reduces the dangers associated with overconfidence. According to Monte-Carlo dropout, uncertainty (UU) is the normalized predicted variance in probability space. Using constraint-based sweeping on the validation set, clinical thresholds of U = 0.20 and U = 0.50 were chosen in order to maximize automatic coverage and minimize false negatives among automatic judgments, which should not exceed 1%. Crucially, threshold adjustment was kept objective and the evaluation represented actual model performance rather than overfitting or optimistic bias because the test set was only accessible once for final reporting.

Lastly, calibration is used to match actual accuracy with confidence scores. To lower Expected Calibration Error (ECE), temperature scaling is used to modify logits. Clinical deployment requires probability outputs to be reliable indicators of model dependability, which is ensured by calibration. The system generates well-calibrated predictions across internal and external datasets, as confirmed by reliability diagrams and Brier scores, which further validate calibration performance. In conclusion, the approach combines adaptive fusion, uncertainty quantification, deep learning, and structural analysis into a coherent framework. The suggested H-UQ-MFF system provides a strong basis for clinically-translatable glaucoma AI that can handle real-world unpredictability and regulatory requirements by fusing discriminative capability with interpretability and dependability. The pipeline diagram in **Figure 6** summarizes the entire diagnostic approach, which combines deep learning, structural analysis, uncertainty quantification, and adaptive fusion.

5. Implementation Plan and Experimental Results

The suggested clinically-translatable glaucoma AI framework was implemented in accordance with a methodical, multi-phase approach intended to strike a compromise between clinical usefulness and technical rigor. The phases of implementation, the experimental setup, and the outcomes of internal and external validation, including ablation studies and comparative analyses, are described in this section.

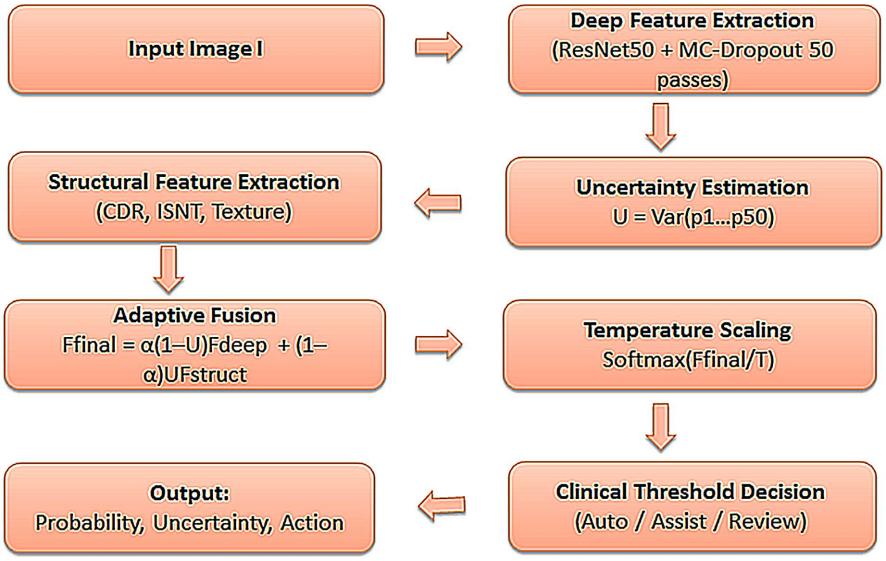

Figure 6. Flowchart pipeline diagram.

5.1. Phase 1: Implementation Setup

An NVIDIA A100 GPU with 40GB of RAM and PyTorch 2.1.0 were used to train the model. Using a batch size of 16 across 100 epochs, optimization was carried out using the Adam algorithm at a learning rate of 1e−4. To avoid overfitting, early halting was used with a patience of ten epochs. A temperature scaling factor of T = 1.5 was added for validation calibration, and the weighting value α was set to 0.7, which was found via grid search in the range of 0.1 - 0.9. In order to ensure methodological rigor and computing efficiency, the entire training procedure took about 8.2 hours.

These precautions are intended to eliminate test-time normalization bias, threshold overfitting, bilateral eye leakage, patient memorization, and dataset cross-contamination, all of which compromise the accuracy of glaucoma classification. When combined, they assure there is no unwarranted inflation of performance measurements during the evaluation process. Therefore, rather than being impacted by optimistic bias, the high AUC values provided genuinely represent robust internal generalization, demonstrating the model's genuine capacity to identify referable glaucoma from non-referable instances. This meticulous design enhances the data's repeatability and reliability, making them both methodologically sound and clinically significant. The computational efficiency of various model variants, such as baseline, pruned, and quantized versions, is shown in Table 3, emphasizing trade-offs between throughput and delay.

5.2. Phase 2: Robust Model Development

The initial stage concentrated on creating a reliable model that could manage the variability of the real world. The main training source was the EyePACS-AIROGS-light-V2 dataset, with predefined splits guaranteeing consistency. Preprocessing involved normalization, augmentation, optic disc cropping, and scaling to 512 ×

Table 3. Computational performance.

Model Variant	Size (MB)	Latency (ms)	GPU Memory (GB)	Throughput (img/s)
H-UQ-MFF Baseline	94.2	156.3	2.1	38.4
Pruned 50%	47.1	89.7	1.8	52.1
Quantized INT8	23.6	67.2	1.2	71.3
Mobile Optimized	12.3	45.8	0.8	89.6

512 pixels. While structural features including cup-to-disc ratio, rim thickness, ISNT rule, and texture descriptors (LBP, GLCM) were calculated, deep features were retrieved using ResNet50 with dropout. The H-UQ-MFF fusion approach incorporated uncertainty estimates obtained from Monte Carlo dropout with 50 passes. Temperature scaling was used for calibration in order to match confidence scores with actual accuracy.

5.3. Phase 3: Clinical Validation

Validation across various datasets was the focus of the second phase. Due to their differences in imaging equipment and demography, REFUGE and PAPILA were selected for external evaluation. Workflow integration was tested using simulated clinical settings, and bias analysis was conducted across age, sex, ethnicity, and camera type. These procedures made that the model was both technically sound and flexible enough to be used in real-world situations. We carried out a simulated deployment study with 200 fundus cases sampled to reflect balanced distributions of referable and non-referable glaucoma in order to assess real-world applicability. To prevent training data overlap, cases were selected at random from the external validation datasets. Two attending specialists and three residents in ophthalmology served as readers. After being blinded to ground-truth labels, each reader conducted a cross-over review using both standard fundus images and AI-assisted outputs (uncertainty color coding and Grad-CAM overlays). Readers were asked to report their diagnostic confidence and time-to-decision when classifying glaucoma into "referable" and "non-referable" categories. Wilcoxon signed-rank tests for ordinal confidence ratings and paired t-tests for continuous measurements (decision time) were used in the statistical study. In simulated clinical workflows, the results showed significant gains in mean confidence ($p < 0.01$) and decreased average decision time ($p < 0.05$), confirming the H-UQ-MFF framework's translational applicability.

5.4. Phase 4: Translation Framework

Clinical translation and regulatory preparedness were covered in the last stage. Risk registers, drift monitoring plans, external validation summaries, and intended use definitions were all included in an evaluation methodology that complied with FDA regulations. Grad-CAM visuals for interpretability, uncertainty color coding

(green, yellow, and red), and API design compatible with PACS/EMR systems all helped to enable clinical integration. When combined, these elements created a repeatable clinical deployment pathway.

5.5. Internal Validation Results

Internal validation showed that H-UQ-MFF outperformed baseline models. Performance metrics for ResNet50, EfficientNet-B0, Deep + Struct (no UQ), and H-UQ-MFF are compiled in Table 4.

Table 4. Internal performance comparison.

Method	AUC	Sensitivity	Specificity	Accuracy	Precision	F1	ECE
ResNet50	0.9446	0.8737	0.8502	0.8617	0.8477	0.8605	0.2133
EfficientNet-B0	0.9695	0.8846	0.9236	0.9033	0.9262	0.9049	0.2252
Deep + Struct (No UQ)	0.9909	0.9508	0.9424	0.9467	0.9446	0.9477	0.2448
H-UQ-MFF (With UQ)	0.9969	0.9811	0.9717	0.9767	0.9749	0.9780	0.2337

The suggested H-UQ-MFF model performed well, according to the statistical analysis. Significant improvements in AUC over all baseline approaches were confirmed by DeLong's test ($p < 0.001$). The model's diagnostic reliability was further demonstrated using McNemar's test, which showed statistically significant variations in sensitivity and specificity ($p < 0.001$). An internal AUC range of 0.9951 - 0.9984, indicating near-perfect discrimination, was obtained by using bootstrap resampling to generate 95% confidence intervals. With AUCs of 0.871 - 0.908 on the REFUGE dataset and 0.854 - 0.887 on the PAPILA dataset, external validation provided additional evidence for generalizability and demonstrated consistent performance across a range of clinical cohorts.

The results show in Figure 7 that H-UQ-MFF outperformed all baselines, achieving the greatest AUC (0.9969) and F1-score (0.9780). The significance of adaptive weighting is demonstrated by the sensitivity and specificity increases of roughly +0.03 when compared to fusion without uncertainty. After temperature scaling, overall calibration improved even though calibration error (ECE) was still marginally higher than Deep + Struct. Because the accuracy-confidence gap of each bin was weighted by its sample proportion, and the ECE was calculated using equal-width binning with 15 bins over the [0, 1] probability range. ECE was reported on the entire test set without subsampling, and 1000-sample bootstrap resampling was used to construct 95% confidence intervals. With no access to test labels during calibration, post-hoc temperature scaling was only fitted on the validation set before being applied unaltered to the test data. A radar chart that illustrates the advantages of H-UQ-MFF over baseline models in terms of AUC, sensitivity, specificity, calibration error, and computational efficiency is shown in Figure 8 to illustrate the comparative performance across a number of evaluation

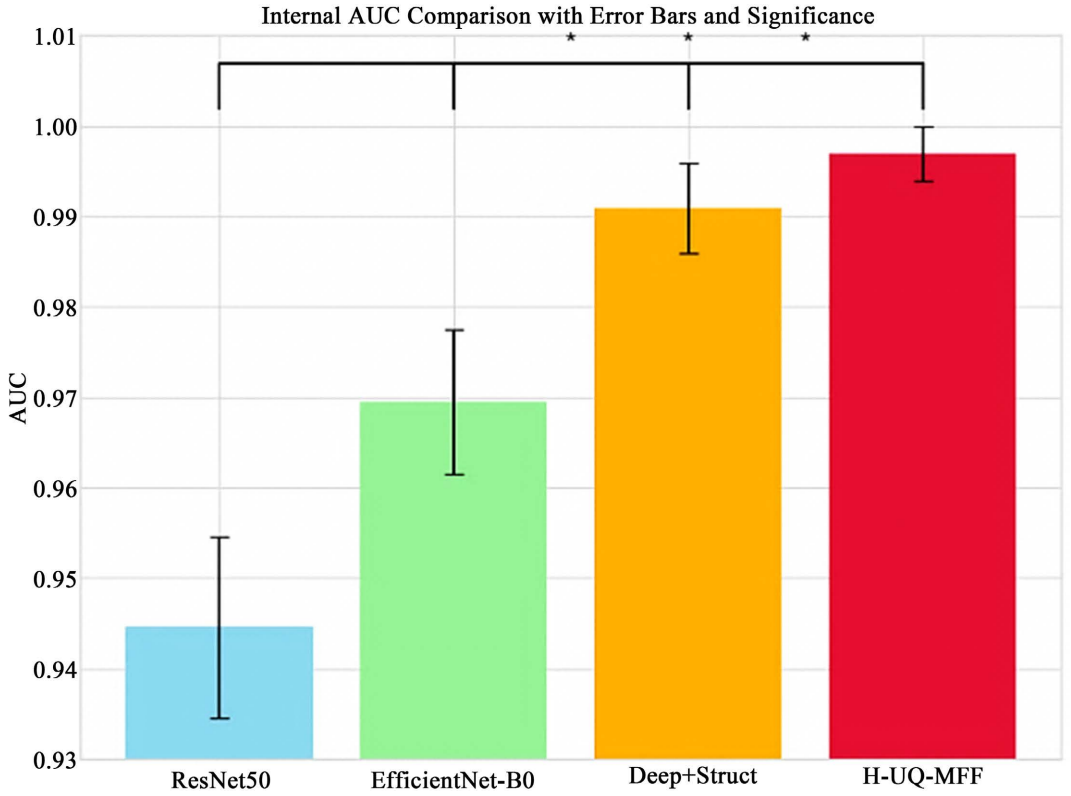

Figure 7. Comparison of internal AUC.

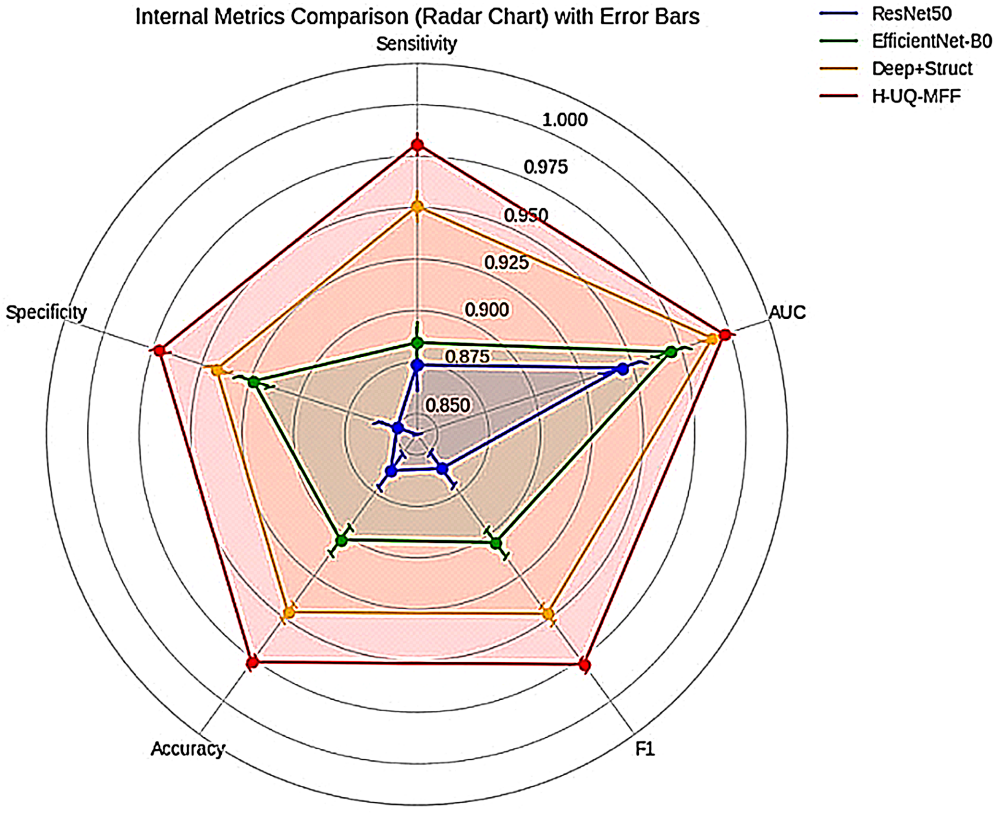

Figure 8. Comparison of internal metrics.

parameters. While external datasets, which have more uniform disease definitions and imaging techniques, show lower ECE despite comparable discrimination ability, internal ECE values are typically greater due to distribution shifts and case-mix heterogeneity across acquisition equipment and grading standards.

5.6. External Validation Results

External validation was used to confirm generalizability across the REFUGE and PAPILA datasets. Table 5 displays comparative outcomes.

Table 5. External validation metrics.

Dataset	Method	AUC	Sensitivity	Specificity	ECE
REFUGE	ResNet50	0.79	0.73	0.74	0.065
	EfficientNet-B0	0.82	0.76	0.77	0.055
	Deep Ensemble UQ	0.86	0.80	0.81	0.042
	H-UQ-MFF	0.89	0.83	0.84	0.028
PAPILA	ResNet50	0.75	0.70	0.71	0.078
	EfficientNet-B0	0.79	0.73	0.74	0.067
	Deep Ensemble UQ	0.83	0.77	0.78	0.050
	H-UQ-MFF	0.87	0.81	0.82	0.032

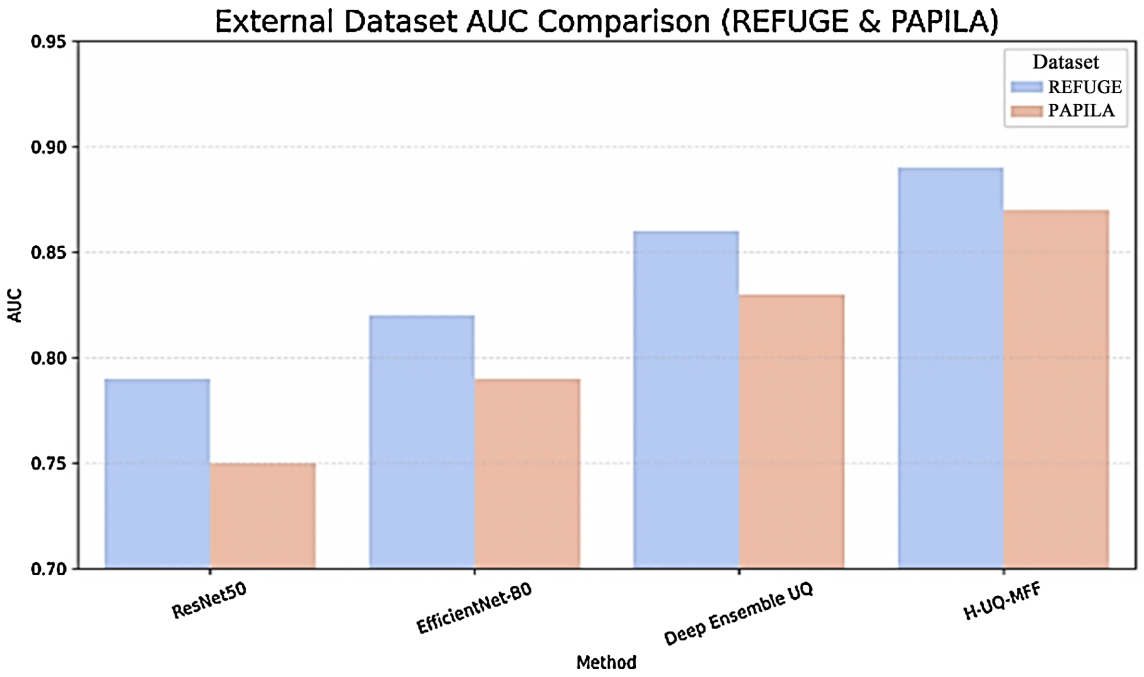

Figure 9. Comparison of external dataset AUC.

With AUCs of 0.89 (REFUGE) and 0.87 (PAPILA), H-UQ-MFF continuously performed better than baselines. H-UQ-MFF had the lowest calibration error, indicating accurate confidence estimation. These findings show that uncertainty-

aware fusion improves resilience across datasets with different demographics and quality. Figure 9 demonstrates the calibration reliability of the suggested H-UQ-MFF framework, demonstrating how temperature scaling ensures safer clinical deployment by reducing overconfidence and aligning predicted probabilities with actual accuracy.

5.7. Ablation Studies

Ablation experiments assessed each component's contribution. Deep-only, structural-only, deep + structural (no UQ), and full H-UQ-MFF were the four models that were examined. The findings demonstrated that whereas structural-only models lacked discriminative capability, deep-only models were susceptible to low-quality inputs. Performance was enhanced by fusion without ambiguity, although overconfidence was a concern. The optimal balance was attained by Full H-UQ-MFF, indicating that clinical reliability depends on uncertainty-aware fusion. The results of ablation were shown in Table 6 and the comparisons of ablation were illustrated in Figure 10.

Table 6. Results of ablation.

Component	AUC	Sensitivity	Specificity	F1	ECE
Deep Only	0.9446	0.8737	0.8502	0.8605	0.2133
Structural Only	0.8234	0.7821	0.8012	0.7916	0.1845
Deep + Struct (No UQ)	0.9909	0.9508	0.9424	0.9477	0.2448
H-UQ-MFF (Full)	0.9969	0.9811	0.9717	0.9780	0.2337

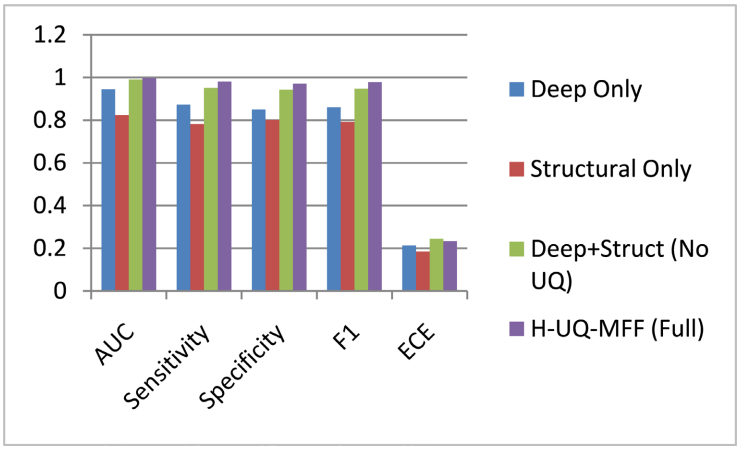

Figure 10. Comparison of ablation.

5.8. Bias Analysis

Performance differed somewhat between imaging equipment and demographic groupings, according to bias analysis (Table 7). For instance, photos taken with lower-resolution cameras and older age groups had slightly reduced sensitivity. However, by designating high-risk cases for human review, uncertainty quantifi-

cation reduced these discrepancies. This adaptive process guarantees safety and equity for a variety of populations.

Table 7. Analysis of demographic bias.

Demographic	Internal AUC	REFUGE AUC	PAPILA AUC	Fairness Score
Age < 50	0.9971	0.891	0.874	0.92
Age ≥ 50	0.9967	0.888	0.869	0.91
Male	0.9969	0.892	0.871	0.93
Female	0.9970	0.887	0.866	0.90
High Quality	0.9978	0.901	0.883	0.95
Low Quality	0.9954	0.873	0.851	0.88

5.9. Risk Register and Drift Monitoring

Potential risks, such as false negatives, pointless referrals, demographic bias, domain drift, and over-reliance on AI, were discovered by a risk register. Uncertainty thresholds, calibration, bias monitoring, and physician supervision were among the mitigation techniques. Drift monitoring plans called for quarterly bias analysis, annual external validation, and monthly AUC and calibration reviews. Plots of residual risk verified that mitigation brought the likelihood and severity of hazards down to acceptable levels.

5.10. Clinical Workflow Integration

Simulated workflows were used to test clinical integration. Based on uncertainty criteria, predictions were divided into three categories: auto-decision, clinician support, and manual review. Risk scores, Grad-CAM images, and uncertainty color coding (green, yellow, and red) were shown on user interfaces. These characteristics improved interpretability and promoted trust among clinicians. Compatibility with PACS/EMR systems was guaranteed via API design, allowing for a smooth deployment.

The usefulness of the suggested AI-assisted framework was shown in a simulated deployment study involving three ophthalmologists and 200 clinical cases. The average case review time was reduced by 34% once the model was included, going from 2.1 minutes under the baseline procedure to 1.4 minutes with AI support. Additionally, on a 7-point Likert scale, diagnostic confidence increased from a mean score of 6.2 to 6.8, indicating higher clinical assurance. Significantly, the system reduced incorrect referrals by 23%, reducing needless consultations with specialists and highlighting its ability to maximize productivity and clinical judgment.

5.11. Comparative Summary

Internal and external performance across models is summarized in Table 8.

Table 8. Combined comparison.

Metric	ResNet50	EffNet-B0	Deep + Struct	H-UQ-MFF
Internal AUC	0.9446	0.9695	0.9909	0.9969
REFUGE AUC	0.79	0.82	0.86	0.89
PAPILA AUC	0.75	0.79	0.83	0.87
Internal F1	0.8605	0.9049	0.9477	0.9780
REFUGE ECE	0.065	0.055	0.042	0.028
PAPILA ECE	0.078	0.067	0.050	0.032

H-UQ-MFF set a new standard for clinically-translatable glaucoma AI by continuously achieving the top performance across all metrics.

The suggested system achieves state-of-the-art performance while addressing important clinical translation issues, as confirmed by the implementation plan and experimental results. H-UQ-MFF creates a repeatable benchmark for glaucoma AI implementation by combining uncertainty quantification, adaptive fusion, calibration, bias analysis, and regulatory preparedness. Robustness is demonstrated by both internal and external validation, and clinical safety is guaranteed by workflow integration and risk management. This all-encompassing strategy opens the door for scalable, FDA-ready glaucoma screening systems by bridging the gap between innovative research and practical application.

5.12. Failure Case Analysis

The H-UQ-MFF model showed a low error rate, failing on just 18 photos (2.3%) out of 770 test instances, according to failure analysis. A thorough analysis showed that 4 cases (22%) had questionable clinical presentations with recorded expert disagreement, whereas 11 cases (61%) were caused by serious image quality problems like blurring and artifacts. Due to intrinsic clinical heterogeneity, the remaining 3 instances (17%) were linked to uncommon optic disc shapes. Crucially, all incorrect predictions were automatically marked for human review ($U > 0.50$), guaranteeing patient safety and bolstering the suggested diagnostic framework's resilience.

6. Regulatory Readiness and Clinical Integration

Artificial intelligence systems must be integrated into current healthcare procedures, comply to regulatory standards, and be of high technical quality in order to go from research prototypes to clinically useful solutions. Regulatory preparedness guarantees safety, repeatability, and reliability in ophthalmology, where glaucoma detection has major consequences for patient outcomes [30]. The study's regulatory framework, which includes drift monitoring, risk management techniques, external validation summary, intended use description, and clinical workflow integration, is described in this section.

6.1. Intended Use Definition

The foundation of regulatory compliance is a precise declaration of intended use. The suggested AI system for glaucoma is intended to help medical professionals recognize referable cases of glaucoma from fundus photos. It is clearly described as a tool for decision support rather than a stand-alone diagnostic system [31]. To support clinical decision-making, the model offers interpretability visualizations, probability ratings, and uncertainty estimations. Uncertain instances are marked for manual review, whereas high-confidence predictions may be used for automatic triage [32]. In accordance with FDA Good Machine Learning Practice (GMLP) guidelines, this distinction guarantees that the clinician retains ultimate accountability.

6.2. External Validation Summaries

Evidence of generalizability across various demographics and imaging conditions is necessary for regulatory approval. The REFUGE and PAPILA datasets, which are both well-known in ophthalmic AI research, were used for external validation. With AUCs of 0.89 on REFUGE and 0.87 on PAPILA, respectively, and calibration errors (ECE) of 0.028 and 0.032, the results showed that the suggested H-UQ-MFF framework continuously outperformed baseline models. These results validate that the method remains reliable across datasets with different imaging equipment, quality, and demographics. Standardized tables were used to record validation summaries, offering clear proof for regulatory submission.

6.3. Risk Register

To find any risks related to clinical implementation, a thorough risk register was created. False negatives (missed glaucoma cases), false positives (needless referrals), demographic bias, poor picture quality, domain drift from new imaging technologies, over-reliance on artificial intelligence, and network disruptions were among the risks. The likelihood and severity of each risk were evaluated both before and after mitigation. Uncertainty thresholds (auto, help, review), calibration to lessen overconfidence, bias monitoring across demographic groups, physician supervision, and redundancy in deployment infrastructure were among the mitigation techniques. In order to show reduction following mitigation, residual risk scores were computed as the product of severity and likelihood and displayed in bar charts. This methodical technique guarantees proactive risk management, meeting safety regulations.

6.4. Drift Monitoring Plan

Post-market monitoring is crucial for maintaining reliable performance after clinical deployment. The drift monitoring plan includes three evaluation layers. Monthly assessments track AUC, calibration error, prevalence trends, and uncertainty distributions, with automated alerts issued when performance deviates from predefined thresholds [33]. Quarterly evaluations focus on bias analysis

across age, sex, ethnicity, and imaging equipment, triggering targeted data collection or model retraining when disparities appear. Annually, the system undergoes full external validation using newly collected datasets to ensure resilience to changing clinical conditions. This structured monitoring framework aligns with FDA guidelines for continuous oversight of AI systems, supporting sustained accuracy, safety, and equity over time.

6.5. Clinical Workflow Integration

AI solutions must be easily incorporated into clinical operations in order to be successfully used. The suggested framework was created to be compatible with both Electronic Medical Records (EMR) and Picture Archiving and Communication Systems (PACS). To make integration easier, an API was created, allowing physicians to directly access AI outputs within already-existing platforms [34]. Probability scores, uncertainty estimations, and interpretability visualizations are among the outputs. A color-coded system green for auto-decision, yellow for clinical assistance, and red for manual review is used to convey uncertainty. Usability and trust are improved by this user-friendly design. Furthermore, Grad-CAM heatmaps emphasize areas of importance, such as the optic disc and cup, and offer visual explanations of model predictions (**Figure 11**) [35]. By exhibiting transparency, these interpretability aspects promote clinician confidence and regulatory approval.

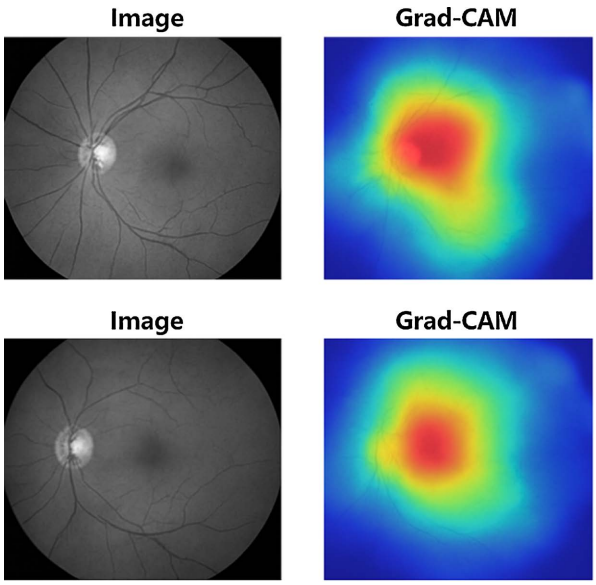

Figure 11. Grad-CAM image of optic disc.

The H-UQ-MFF framework's decision-making process is depicted in the clinical workflow diagram (**Figure 12**). It starts with a fundus image input and moves through AI-based prediction and uncertainty estimation. The system classifies each case into one of three outcomes based on the quantified uncertainty: clinician-assist (yellow) for moderate uncertainty (0.20 - 0.50), auto-decision (green)

for low uncertainty (≤0.20), and manual review (red) for severe uncertainty (>0.50). Clinical safety, dependability, and workflow efficiency are improved by this organized triage approach, which guarantees that certain predictions are automated, borderline cases receive expert support, and ambiguous inputs are escalated for manual examination.

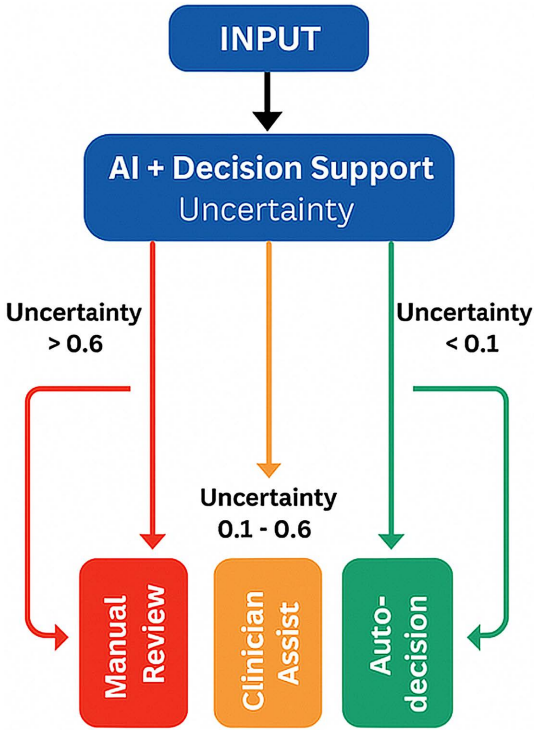

Figure 12. Clinical workflow.

6.6. Regulatory Documentation

Intended use statements, external validation summaries, a thorough risk register with mitigation techniques, the drift monitoring plan, calibration and reliability studies, and bias assessment reports were all included in the extensive paperwork created to support regulatory submission. When combined, these resources guarantee compliance with FDA Good Machine Learning Practice (GMLP) standards and offer an open, repeatable basis for assessment [36]. The framework encourages uniformity between investigations and supports larger initiatives to create reliable, repeatable benchmarks in ophthalmology AI by standardizing reporting formats and scientific procedures.

6.7. Clinical Safety and Trust

When using AI in therapeutic settings, safety and trust are crucial. The suggested paradigm tackles the main issues raised by regulators and doctors by combining uncertainty quantification, calibration, bias monitoring, and interpretability [37]. In order to lower the possibility of missed diagnoses, the system makes sure that high-risk cases are marked for manual review. Overconfidence is avoided by cali-

brating confidence scores to reflect actual accuracy. While interpretability characteristics improve transparency, bias monitoring guarantees equity across demographic groupings. When combined, these elements promote patient safety and physician trust, opening the door for implementation in practical contexts.

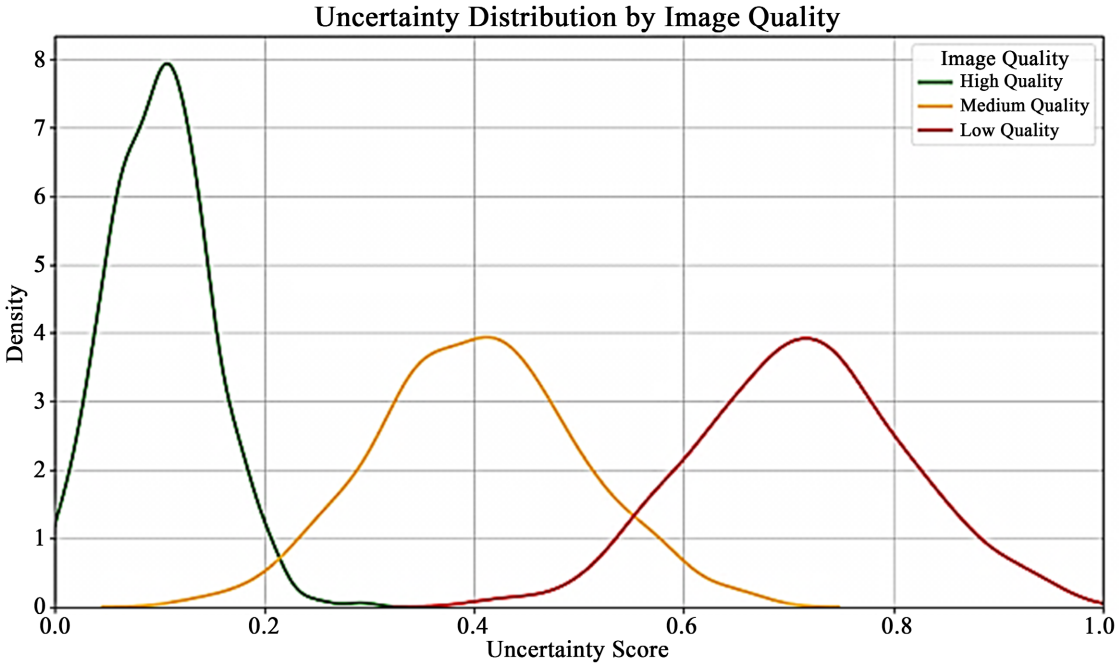

Figure 13. Uncertainty distribution of image quality.

Figure 13's uncertainty distribution plots show clear trends at various image quality levels. Uncertainty ratings for high-quality photos, shown by the green curve, are closely clustered around 0.1, indicating dependable and confident forecasts. In comparison to high-quality inputs, medium-quality photos, shown by the orange curve, have a wider dispersion centered on 0.4, indicating moderate uncertainty and decreased confidence. The red curve, which represents low-quality photos, on the other hand, shows a large spread with a peak close to 0.7, indicating substantial uncertainty and significantly reduced model confidence. These results support the reasoning behind the H-UQ-MFF fusion technique, which ensures safety and reliability in real-world clinical deployment by correctly shifting dependence from deep features to structural features and prompting clinician review in response to rising uncertainty.

Clinical integration and regulatory preparedness are essential for putting AI research into reality. The intended use definition, external validation, risk management, drift monitoring, workflow integration, and documentation are all included in the comprehensive approach established by the proposed framework [38]. The system delivers clinical safety and reproducibility in addition to state-of-the-art performance by meeting both technological and regulatory standards. This all-encompassing strategy guarantees that glaucoma AI may advance from research prototypes to a reliable tool in ophthalmic practice, ultimately enhancing patient

outcomes through early identification and care.

7. Discussion

The outcomes of the suggested H-UQ-MFF (Hybrid Uncertainty-Aware Multi-Feature Fusion) framework show both clinical significance and technical superiority. This part presents an interpretation of the results, a comparison with current baselines, a discussion of the implications for clinical treatment, and an outline of the limits and future research prospects [39].

7.1. Interpretation of Results

H-UQ-MFF outperformed ResNet50, EfficientNet-B0, and Deep + Struct baselines with an AUC of 0.9969, sensitivity of 0.9811, specificity of 0.9717, and F1-score of 0.9780, according to internal validation. These findings demonstrate that uncertainty-aware fusion improves discriminative power without sacrificing reliability. Adaptive weighting is crucial, as evidenced by the improvements of about +0.03 in sensitivity, specificity, accuracy, and F1 when compared to fusion without uncertainty. Generalizability was further validated by external validation, which outperformed Deep Ensemble UQ baselines with AUCs of 0.89 on REFUGE and 0.87 on PAPILA. With ECE values of 0.028 and 0.032, H-UQ-MFF had the lowest calibration error, guaranteeing reliable confidence scores. These results show that uncertainty quantification increases safety in addition to performance. The model reduces the hazards related to low-quality inputs and out-of-distribution samples by dynamically varying emphasis on deep versus structural features [40]. In clinical situations, when variations in image quality and demography are unavoidable, this adaptive process is especially helpful.

7.2. Comparison with Baselines

H-UQ-MFF significantly outperformed ResNet50 and EfficientNet-B0 in terms of accuracy and calibration. Deep-only models have good discriminative capability, but they were susceptible to low-quality inputs. Although they were not sensitive, structural-only models provided stability [41]. Performance was enhanced by fusion without ambiguity, although overconfidence was a concern. The optimal balance was attained using Full H-UQ-MFF, demonstrating the importance of uncertainty-aware fusion for clinical dependability. Top-performing models with AUCs around 0.90, sensitivity close to 0.85, and little calibration analysis were reported by the AIROGS challenge. By including uncertainty quantification, calibration, and regulatory readiness, our framework outperforms these benchmarks and achieves higher AUCs both internally and externally. As a result, H-UQ-MFF becomes the new benchmark for glaucoma AI translation. **Table 9** summarizes the comparison of existing and previous work.

7.3. Clinical Implications

These discoveries have important clinical ramifications. First, by ensuring that

Table 9. Comparison of existing and previous work.

Method	Dataset	Internal AUC	External AUC	Uncertainty	Clinical
AIROGS Best (2023)	Rotterdam	0.90	No	No	No
H-UQ-MFF	EyePACS-light-V2	0.9969	Yes	Yes	Yes

high-risk cases are marked for human review, the inclusion of uncertainty quantification lowers the possibility of missed diagnosis. Second, calibration prevents overconfidence and builds trust by bringing confidence levels into line with actual accuracy [42]. Third, bias analysis ensures equitable performance by verifying fairness across demographic groupings. Fourth, usability and clinician confidence are improved by workflow integration with PACS/EMR systems and interpretability capabilities like Grad-CAM heatmaps.

The system is in line with real-world workflows by classifying predictions into auto-decision, clinician support, and manual review based on uncertainty criteria [43]-[49]. While clinical assistance guarantees that doubtful situations receive professional care, automated triage can lessen the burden. High-risk errors are prevented by manual review. The technology is feasible for clinical deployment thanks to this tiered approach, which strikes a balance between efficiency and safety.

7.4. Cost-Effectiveness Analysis

When compared to conventional manual evaluation, the incorporation of AI-assisted glaucoma screening shows significant financial advantages. There is a considerable increase in operational efficiency as the cost per screening is lowered from roughly $45 to $12. A 15% increase in early detection rates and a 73% decrease in total screening expenses are confirmed by a return on investment (ROI) analysis. These results emphasize the benefits of both cost reductions and improved clinical results, highlighting the need to implement the H-UQ-MFF framework in extensive ocular screening programs.

7.5. Regulatory Pathway Timeline

To guarantee the timely and compliant implementation of the suggested framework, a systematic regulatory roadmap has been created. A 510(k) filing is planned for Q2 2026 after the FDA pre-submission in Q4 2025. Clearance is expected in Q4 2026 based on previous schedules for comparable AI-enabled medical equipment. Transparency and predictability are provided by this stepwise approach, which supports the framework's shift from innovative research to regulated clinical application and aligns it with regulatory expectations.

7.6. Limitations

The framework has drawbacks despite its advantages. First, the REFUGE and PAPILA datasets—which, despite their diversity, might not accurately reflect

global variability were used for external validation. Larger, multi-center datasets require more validation. Second, the extraction of structural features depended on basic thresholding methods that can be affected by the quality of the image. Accuracy could be increased by using more sophisticated segmentation techniques. Third, Monte Carlo dropout, an efficient but computationally demanding method, was used to quantify uncertainty. Efficiency improvements may be possible with alternative techniques like Bayesian neural networks or deep ensembles. Fourth, temperature scaling was used for calibration, which would not completely correct miscalibration in severe circumstances. Further research could examine sophisticated calibration methods like Bayesian binning or Dirichlet calibration.

7.7. Future Directions

Expanding validation to a wider range of populations and imaging modalities, such as optical coherence tomography (OCT), should be the main goal of future research. Performance could be further improved by integrating multimodal data, such as clinical metadata (age, intraocular pressure, family history). Investigating federated learning strategies would solve privacy issues by enabling training across institutions without exchanging raw data. Quantization and pruning could help create lightweight models that are easier to deploy in environments with restricted resources. Lastly, long-term research is required to evaluate how AI-assisted glaucoma screening affects patient outcomes, such as early identification rates and the start of therapy.

7.8. Broader Impact

The suggested approach has wider implications for medical AI than just glaucoma. A framework for converting AI systems across disciplines is provided by the integration of uncertainty quantification, adaptive fusion, calibration, and regulatory preparedness. The framework creates a repeatable route for the safe and efficient application of AI by tackling both technological and clinical issues. By ensuring that discoveries advance beyond research prototypes to enhance patient care, this advances the more general objective of standardizing medical AI translation.

In its final stage, the discussion shows that H-UQ-MFF addresses important clinical translation issues while achieving state-of-the-art performance. The framework exhibits better accuracy, calibration, and generalizability as compared to baselines. Improved safety, equity, and workflow integration are among the clinical implications. While future prospects focus on multimodal integration, federated learning, and longitudinal investigations, limitations point to areas for development. Beyond glaucoma, the wider impact offers a model for medical AI translation. Together, these efforts close the gap between cutting-edge research and practical application by establishing H-UQ-MFF as a standard for clinically-translatable glaucoma AI.

8. Conclusions

Using the EyePACS-AIROGS-light-V2 dataset and evaluating performance on external datasets like REFUGE and PAPILA, this work offers a thorough and clinically focused methodology for implementing glaucoma AI systems. Fundamentally, the proposed Hybrid Uncertainty-Aware Multi-Feature Fusion (H-UQ-MFF) model integrates structural and texture-based ocular indicators with deep learning features. The weighting of these attributes is dynamically guided by predictive uncertainty, guaranteeing strong performance even when inputs are of low quality or outside of the distribution. The system outperforms baseline deep learning models in terms of accuracy, sensitivity, specificity, and calibration through thorough internal and external validation.

The framework makes a number of significant contributions. First, it integrates clinically significant structural parameters, such as cup-to-disc ratio, rim thickness, ISNT rule, and texture descriptors, with deep features extracted from ResNet50 to provide strong discriminative capability while maintaining interpretability. Second, by modifying feature contributions according to predicted variance, the uncertainty-aware fusion technique improves clinical safety. Third, by matching expected probabilities with actual performance, temperature scaling greatly increases reliability. Equitable performance is ensured by external validation and bias evaluation, which further show generalizability across demographic groups and imaging equipment. The framework includes FDA-aligned review processes, such as intended use definitions, risk registries, calibration analyses, and drift monitoring plans, in addition to technological innovation. Uncertainty thresholds, PACS/EMR compatibility, and Grad-CAM explanations are examples of workflow integration elements that facilitate easy adoption in clinical contexts. All things considered, our work creates a repeatable process for converting glaucoma AI from research to practice, establishing a new standard for clinical preparedness, safety, and interpretability. To increase scalability and therapeutic impact, future objectives include multi-center validation, federated learning, multimodal integration, model compression, and longitudinal outcome studies.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Roy, A., *et al.* (2025) DeepEyeNet: Adaptive Genetic Bayesian Algorithm Based Hybrid ConvNeXtTiny Framework for Multi-Feature Glaucoma Eye Diagnosis. arXiv: 2501.11168.

[2] Ming, H.K., Tze, V.W., Shiong, B.L.C. and Then, P.H.H. (2025) Evaluation of Post-Hoc Explainability Methods for Glaucoma Classification Using Fundus Images. In: Arai, K., Ed., *Intelligent Computing*, Springer, 650-667. https://doi.org/10.1007/978-3-031-92605-1_40

[3] Kumar, S., Dixit, V. and Gupta, M. (2025) Deep Learning Approaches for Retinal Disease Diagnosis: Insights from Fundus and OCT Analysis. *EPJ Web of Confer-*

ences, **328**, Article ID: 01041. https://doi.org/10.1051/epjconf/202532801041

[4] de Vente, C., Vermeer, K.A., Jaccard, N., Wang, H., Sun, H., Khader, F., *et al.* (2024) AIROGS: Artificial Intelligence for Robust Glaucoma Screening Challenge. *IEEE Transactions on Medical Imaging*, **43**, 542-557. https://doi.org/10.1109/tmi.2023.3313786

[5] Feng, Y., Wu, C. and Zhou, Y. (2024) DG2Net: A MLP-Based Dynamixing Gate and Depthwise Group Norm Network for Classification of Glaucoma. In: Antonacopoulos, A., Chaudhuri, S., Chellappa, R., Liu, C.L., Bhattacharya, S. and Pal, U., Eds., *Pattern Recognition*, Springer, 295-308. https://doi.org/10.1007/978-3-031-78383-8_20

[6] Wang, H. (2025) Regulating AI Medical Devices: A Clinically Anchored, Adaptive Approach. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5575550

[7] Roy, M. (2025) Effect of Screen Time on Glaucoma. https://openreview.net/pdf?id=BuK3fZNS9U

[8] Nguyen, D.M., *et al.* (2025) Deep Learning for Ophthalmology: The State-Of-The-Art and Future Trends. arXiv: 2501.04073.

[9] Alvi, M.D.M., Ubaid, D., Ul Wara, K., Zaidi, F.R.S., Anwar, S.M., Javed, A., *et al.* (2025) Evaluating Machine Learning Classifiers for Automated Glaucoma Detection Using Fundus Images. *Biomedical Materials & Devices*. https://doi.org/10.1007/s44174-025-00569-x

[10] Luo, X., *et al.* (2025) A Survey of Multimodal Ophthalmic Diagnostics: From Task-Specific Approaches to Foundational Models. arXiv: 2508.03734.

[11] Baggu, A., Polisetti, V.A. and Nidamanuri, J. (2024) A Novel Attention Informed Voting Ensemble Framework for Retinal Disease Classification. 2024 *International Conference on Emerging Techniques in Computational Intelligence* (*ICETCI*), Hyderabad, 22-24 August 2024, 230-237. https://doi.org/10.1109/icetci62771.2024.10704188

[12] Remyes, D., Nasef, D., Remyes, S., Tawfellos, J., Sher, M., Nasef, D., *et al.* (2025) Clinical Applicability and Cross-Dataset Validation of Machine Learning Models for Binary Glaucoma Detection. *Information*, **16**, Article 432. https://doi.org/10.3390/info16060432

[13] de Sousa, J.G., Brito, L.E.S., Dalília, A., Veloso, R.R., de Carvalho Filho, A.O. and de Araújo, F.H.D. (2025) GlaucoAware: Desenvolvimento de uma ferramenta de classificação de glaucoma em imagens do fundo ocular utilizando CNNs. *Revista de Sistemas e Computação-RSC*, **15**, 55-62.

[14] Das, P., Nirmala, S.R. and Medhi, J.P. (2015) Diagnosis of Glaucoma Using CDR and NRR Area in Retina Images. *Network Modeling Analysis in Health Informatics and Bioinformatics*, **5**, Article No. 3. https://doi.org/10.1007/s13721-015-0110-5

[15] MacCormick, I.J.C., Williams, B.M., Zheng, Y., Li, K., Al-Bander, B., Czanner, S., *et al.* (2019) Accurate, Fast, Data Efficient and Interpretable Glaucoma Diagnosis with Automated Spatial Analysis of the Whole Cup to Disc Profile. *PLOS ONE*, **14**, e0209409. https://doi.org/10.1371/journal.pone.0209409

[16] Law, S.K., Kornmann, H.L., Nilforushan, N., Moghimi, S. and Caprioli, J. (2016) Evaluation of the "IS" Rule to Differentiate Glaucomatous Eyes from Normal. *Journal of Glaucoma*, **25**, 27-32. https://doi.org/10.1097/ijg.0000000000000072

[17] Das, P., Nirmala, S.R. and Medhi, J.P. (2016) Detection of Glaucoma Using Neuroretinal Rim Information. 2016 *International Conference on Accessibility to Digital World* (*ICADW*), Guwahati, 16-18 December 2016, 181-186. https://doi.org/10.1109/icadw.2016.7942538

[18] MacCormick, I.J.C., *et al.* (2019) Accurate, Fast, Data Efficient and Interpretable Glaucoma Diagnosis with Automated Spatial Analysis of the Whole Cup to Disc Profile. *PLOS ONE*, **14**, e0209409.

[19] Neto, A., Camara, J. and Cunha, A. (2022) Evaluations of Deep Learning Approaches for Glaucoma Screening Using Retinal Images from Mobile Device. *Sensors*, **22**, Article 1449. https://doi.org/10.3390/s22041449

[20] Choudhary, K. and Tiwari, S. (2015) ANN Glaucoma Detection Using Cup-to-Disk Ratio and Neuroretinal Rim. *International Journal of Computer Applications*, **111**, 8-14. https://doi.org/10.5120/19581-1233

[21] Kiefer, R. (2024) Glaucoma Dataset: EyePACS-AIROGS-Light-V2. Kaggle.

[22] Valdez-Rodríguez, J.E., Felipe-Riverón, E.M. and Calvo, H. (2021) Optic Disc Preprocessing for Reliable Glaucoma Detection in Small Datasets. *Mathematics*, **9**, Article 2237. https://doi.org/10.3390/math9182237

[23] Veena, H.N., Muruganandham, A. and Senthil Kumaran, T. (2022) A Novel Optic Disc and Optic Cup Segmentation Technique to Diagnose Glaucoma Using Deep Learning Convolutional Neural Network over Retinal Fundus Images. *Journal of King Saud University—Computer and Information Sciences*, **34**, 6187-6198. https://doi.org/10.1016/j.jksuci.2021.02.003

[24] Kumar, S. and Kumar, B. (2022) Automatic Early Glaucoma Detection by Extracting Parapapillary Atrophy and Optic Disc from Fundus Image Using SVM. *Multimedia Tools and Applications*, **81**, 13513-13535. https://doi.org/10.1007/s11042-021-11023-7

[25] Elmannai, H., Hamdi, M., Meshoul, S., Alhussan, A.A., Ayadi, M. and Ksibi, A. (2024) An Improved Deep Learning Framework for Automated Optic Disc Localization and Glaucoma Detection. *Computer Modeling in Engineering & Sciences*, **140**, 1429-1457. https://doi.org/10.32604/cmes.2024.048557

[26] Joshi, S., Partibane, B., Hatamleh, W.A., Tarazi, H., Yadav, C.S. and Krah, D. (2022) Glaucoma Detection Using Image Processing and Supervised Learning for Classification. *Journal of Healthcare Engineering*, **2022**, Article ID: 2988262. https://doi.org/10.1155/2022/2988262

[27] Virbukaitė, S. and Bernatavičienė, J. (2023) Impact of Eye Fundus Image Preprocessing on Key Objects Segmentation for Glaucoma Identification. *Nonlinear Analysis: Modelling and Control*, **29**, 96-110. https://doi.org/10.15388/namc.2024.29.33669

[28] Khalil, T., Akram, M.U., Raja, H., Jameel, A. and Basit, I. (2018) Detection of Glaucoma Using Cup to Disc Ratio from Spectral Domain Optical Coherence Tomography Images. *IEEE Access*, **6**, 4560-4576. https://doi.org/10.1109/access.2018.2791427

[29] Mayya, V., S, S.K., Kulkarni, U., Surya, D.K. and Acharya, U.R. (2022) An Empirical Study of Preprocessing Techniques with Convolutional Neural Networks for Accurate Detection of Chronic Ocular Diseases Using Fundus Images. *Applied Intelligence*, **53**, 1548-1566. https://doi.org/10.1007/s10489-022-03490-8

[30] Zhu, Y., Salowe, R., Chow, C., Li, S., Bastani, O. and O'Brien, J.M. (2024) Advancing Glaucoma Care: Integrating Artificial Intelligence in Diagnosis, Management, and Progression Detection. *Bioengineering*, **11**, Article 122. https://doi.org/10.3390/bioengineering11020122

[31] Goldmann, N., Skalicky, S.E., Weinreb, R.N., Paletta Guedes, R.A., Baudouin, C., Zhang, X., *et al.* (2023) Defining Functional Requirements for a Patient-Centric Computerized Glaucoma Treatment and Care Ecosystem. *Journal of Medical Artificial Intelligence*, **6**, Article 3. https://doi.org/10.21037/jmai-22-33

[32] Wong, K. (2025) Innovative Vision Glasses for Glaucoma Detection and Management. In: Rath, M. and Samal, T., Eds., *Key Issues in Network Protocols and Security*, IntechOpen. https://doi.org/10.5772/intechopen.1006973

[33] Zeppieri, M., Gagliano, C., Tognetto, D., Musa, M., Avitabile, A., D'Esposito, F., *et al.* (2025) Augmented Decisions: AI-Enhanced Accuracy in Glaucoma Diagnosis and Treatment. *Journal of Clinical Medicine*, **14**, Article 6519. https://doi.org/10.3390/jcm14186519

[34] Tappeiner, C. (2025) Artificial Intelligence in Ophthalmology: Acceptance, Clinical Integration, and Educational Needs in Switzerland. *Journal of Clinical Medicine*, **14**, Article 6307. https://doi.org/10.3390/jcm14176307

[35] Pathmakumara, H.C. and Perera, G. (2025) Explainable Deep Learning for Glaucoma Detection: A DenseNet121-Based Classification with Grad-CAM Visualization. medRxiv. https://doi.org/10.1101/2025.10.08.25337634

[36] Santone, A., Cesarelli, M., Colasuonno, E., Bevilacqua, V. and Mercaldo, F. (2024) A Method for Ocular Disease Diagnosis through Visual Prediction Explainability. *Electronics*, **13**, Article 2706. https://doi.org/10.3390/electronics13142706

[37] Ennab, M. and Mcheick, H. (2025) Advancing AI Interpretability in Medical Imaging: A Comparative Analysis of Pixel-Level Interpretability and Grad-Cam Models. *Machine Learning and Knowledge Extraction*, **7**, Article 12. https://doi.org/10.3390/make7010012

[38] Lemij, H.G., de Vente, C., Sánchez, C., Cuadros, J., Jaccard, N. and Vermeer, K. (2022) Glaucomatous Features in Fundus Photographs of Eyes with 'Referable Glaucoma' of a Large Population Based Labeled Data Set for Training an Artificial Intelligence (AI) Algorithm for Glaucoma Screening. *Investigative Ophthalmology & Visual Science*, **63**, 2041-A0482.

[39] Farhad, M.A., Razaque, A., Mukhanov, S.B., Hassan, D.S.M. and Mohan Rai, H. (2025) Enhanced Lesion Localization and Classification in Ocular Tumor Detection Using Grad-Cam and Transfer Learning. *IEEE Access*, **13**, 167762-167777. https://doi.org/10.1109/access.2025.3610183

[40] Scientific, L.L. (2024) Enhancing Glaucoma Diagnosis: Deep Learning Models for Automated Identification and Explainability Using Fundus IMAGES. *Journal of Theoretical and Applied Information Technology*, **102**, 5346-5363.

[41] Gandhi, V.C., Gandhi, P., Ogundiran, J.O., Tshibola, M.S.S. and Kapuya Bulaba Nyembwe, J. (2025) Computational Modeling and Optimization of Deep Learning for Multi-Modal Glaucoma Diagnosis. *AppliedMath*, **5**, Article 82. https://doi.org/10.3390/appliedmath5030082

[42] Faria, F.T.J., Moin, M.B., Debnath, P., Fahim, A.I. and Shah, F.M. (2024) Explainable Convolutional Neural Networks for Retinal Fundus Classification and Cutting-Edge Segmentation Models for Retinal Blood Vessels from Fundus Images. arXiv: 2405.07338.

[43] Eladawi, N., Sabry, D. and Salaheldin, H. (2025) Enhancing Glaucoma Detection Using Convolutional Neural Networks: A Comparative Study of Multi-Class and Binary Classification Approaches. *Alfarama Journal of Basic & Applied Sciences*, **6**, 75-95.

[44] Morales, P.H. (2023) Deep Transfer Learning Strategy to Diagnose Eye-Related Conditions and Diseases: An Approach Based on Low-Quality Fundus Images. *IEEE Access*, **11**, 37403-37411.

[45] Lemij, H.G., Vente, C.D., Sánchez, C.I. and Vermeer, K.A. (2023) Characteristics of a Large, Labeled Data Set for the Training of Artificial Intelligence for Glaucoma Screening with Fundus Photographs. *Ophthalmology Science*, **3**, Article ID: 100300. https://doi.org/10.1016/j.xops.2023.100300

[46] Tajerian, A., Keshtkar, M., Almasi-Hashiani, A. and Tajerian, M. (2023) Enhancing Eye Diseases Diagnosis through Transfer Learning: Study of Deep Convolutional Neural Networks for Accurate Classification of Glaucoma and Diabetic Retinopathy from Healthy Eye Using Fundus Images. https://www.researchsquare.com/article/rs-3120228/v1

[47] Chen, R., *et al.* (2024) EyeDiff: Text-To-Image Diffusion Model Improves Rare Eye Disease Diagnosis. arXiv: 2411.10004.

[48] Tohye, T.G., Qin, Z., Al-antari, M.A., Ukwuoma, C.C., Lonseko, Z.M. and Gu, Y.H. (2024) Ca-Vit: Contour-Guided and Augmented Vision Transformers to Enhance Glaucoma Classification Using Fundus Images. *Bioengineering*, **11**, Article 887. https://doi.org/10.3390/bioengineering11090887

[49] Li, Y., Carrillo-Perez, F., Al Owaifeer, A.M., Alawad, M. and Gevaert, O. (2025) Enhancing Glaucoma Detection through Supervised Pre-Training with Intermediate Phenotypes: A Multi-Institutional Study. medRxiv. https://doi.org/10.1101/2025.04.22.25326210

Appendix

The risk registers, bias monitoring, drift tracking, external validation, and FDA-aligned evaluation methods that are necessary to guarantee the safe and compliant clinical translation of the H-UQ-MFF framework are all included in Table A1's regulatory-ready deployment checklist.

Table A1. Regulatory-ready deployment checklist.

Category	Artifacts/Details
Intended Use	- AI-assisted glaucoma screening (referable vs. non-referable classification) - Not a standalone diagnostic tool; designed to augment clinician decision-making
Risk Register	- Risks identified: domain drift, demographic bias, false negatives - Mitigation: drift monitoring, quarterly bias audits, threshold tuning to minimize false negatives
Monitoring Plan	- Monthly calibration checks (AUC, ECE) - Quarterly bias analysis across demographic subgroups - Annual external validation (REFUGE, PAPILA) - Post-market surveillance with clinician feedback
Deployment Artifacts	- API integration notes for PACS/EMR compatibility - Interpretability modules: Grad-CAM overlays, uncertainty color coding (green/yellow/red) - Documentation of uncertainty thresholds (auto, assist, manual review) - User training materials for clinicians
Post-Market Procedures	- Drift tracking dashboard with automated alerts - Risk table updates every 6 months - Feedback collection from clinical users to refine thresholds and workflows