A Web-Based Clinical Decision-Support Dashboard for Glaucoma Screening from Retinal Fundus Images: A Two-Phase VGG16 Deep Learning Pipeline with Grad-CAM Interpretability for Ophthalmology Practice

Shafiya Qadeer Memon[1]*, Rafay Shakeel[1], Amirita Dewani[1], Mehran Memon[2], Sachal Sahito[1], Areeba Syed[1],
[1]Department of Software Engineering, Mehran University of Engineering and Technology, Jamshoro 76062, Sindh, Pakistan
[2]Sindh Institute of Ophthalmology and Visual Sciences (SIOVS), Hyderabad, Sindh, Pakistan
***Correspondence**: shafiya.memon@faculty.muet.edu.pk

Citation | Memon. S. Q, Shakeel. R, Amrita Dewani, Memon. M Sahito. S, Syed. A, "A Web-Based Clinical Decision-Support Dashboard for Glaucoma Screening from Retinal Fundus Images: A Two-Phase VGG16 Deep Learning Pipeline with Grad-CAM Interpretability for Ophthalmology Practice", IJIST, Vol. 8 Issue. 6 pp 2452-2467, September 2026
DOI: https://doi.org/10.33411/IJIST/2015
Received | Aug 16, 2026 **Revised |** Sep 04, 2026 **Accepted |** Sep 07, 2026 **Published |** Sep 08, 2026.

Objective: To develop and evaluate a two-phase transfer-learning pipeline using a VGG16 convolutional neural network for automated binary classification of glaucoma from retinal fundus images, designed for deployment in resource-constrained South Asian screening contexts.

Methods: A retrospective diagnostic-accuracy study used 705 retinal fundus images from the public ACRIMA dataset. The model employed a VGG16 backbone pre-trained on ImageNet with a custom classification head (global average pooling, two dense ReLU layers with dropout, sigmoid output). Phase-1 trained the head with the backbone frozen (Adam, lr $1 \times 10-4$, 50 epochs). Phase-2 unfroze the last four convolutional block and fine-tuned at lr $1 \times 10-5$ for 25 epochs. Online augmentation, inverse-frequency class weighting, early stopping and learning-rate reduction on plateau were applied. A held-out test set (n = 140, balanced 70/70) was used for final evaluation. A FastAPI inference service and a Next.js clinical dashboard were developed for clinician-facing deployment.

Results: The fine-tuned network achieved accuracy 88.57% (95% CI: 82.2–92.8%), sensitivity 92.86% (95% CI: 84.3–96.9%), specificity 84.29% (95% CI: 74.0–91.0%), precision 85.53% (95% CI: 75.9–91.7%), F1-score 89.04%, and AUC-ROC 0.954 (95% CI: 0.918–0.990). Phase-2 fine-tuning raised sensitivity by 7.15 percentage points over Phase-1 (85.71% → 92.86%) and reduced the false-negative rate from 14.3% to 7.1%. Grad-CAM heatmaps consistently localized activation over the optic disc and peripapillary region.

Conclusion: A computationally tractable, interpretable VGG16 pipeline reached screening-level sensitivity with a clinician-facing dashboard, providing a reproducible foundation for region-adapted glaucoma screening pending prospective validation on a Pakistani cohort.

Keywords: Glaucoma, Deep learning, Convolutional neural network, Transfer learning, Fundus imaging, Grad-CAM

Introduction:

Glaucoma is a disease of the optic nerve (neuropathy) and is the most common cause of irreversible blindness in the world, and is expected to impact over 111 million people by 2040 [1]. The disease may be quite insidious, meaning that it can only be diagnosed once significant vision loss has taken place; therefore, timely diagnosis is the main variable that can be influenced to impact the disease outcome. Glaucoma is projected to affect approximately 1.8 million people in Pakistan, with many people likely undiagnosed due to a lack of access to glaucoma specialists in rural areas [2][3]. Current methods of glaucoma assessment include parameter visual field testing, intraocular pressure (IOP) tonometry and expert evaluation of the optic-nerve-head (ONH) on either a fundus photograph or optical coherence tomography (OCT). These techniques are specialist techniques that are time-consuming and have an inter-observer variability [4]. In low and middle income countries, the majority of the world, they are farthest behind in numbers of available trained ophthalmologists, the level they are confronted with and the throughput that is able to be achieved by implementing structured screening and presenting patients for treatment in a timely manner [5]. Image analysis is thus rising to a key adjunct to manual analysis, particularly by using convolutional neural networks (CNNs), which are seen as a viable alternative to manual triage procedures and could be incorporated into primary-care or telemedicine processes [6]. Figure 1 displays a normal retinal image in which the optic disc, the primary anatomical feature of interest for glaucoma classification, can be seen in the middle of the image. Raw input without pre-processing is also shown surrounding the retinal vasculature and the macula.

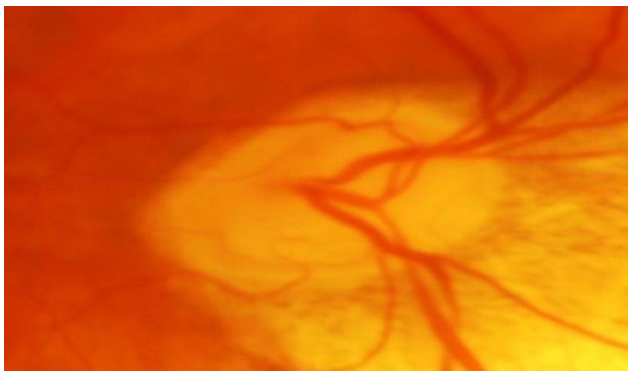

Figure 1. Representative color fundus photograph used for deep-learning-based glaucoma classification. The optic disc the primary anatomical target for automated analysis is visible at the center of the field.

Related Work:

Deep learning for glaucoma detection has matured rapidly since the seminal demonstration by Li and colleagues that an Inception-v3 network could achieve an AUC of 0.986 on 48,116 labelled fundus images [7]. Later on, [8] compared and evaluated 5 CNN models pre-trained with the ImageNet dataset on five public databases, publishing the ACRIMA database utilized in their experiments and achieving a best AUC score of 0.961 on this data with Xception. The REFUGE challenge established multi-center evaluation on a set of 1,200 images and offered top ranking teams better performance than two different glaucoma experts when performing binary classification of glaucoma [9]. Disc-aware ensembles (in which the four-stream fusion architecture reported in [10] achieved a higher AUC than did single baseline) have been researched alongside two-stage architectures comprising RCNN based disc localization, CNN based disc classification (as per Bajwa et al. [11] who reported an AUC of 0.874 on ORIGA), and custom-depth networks trained from scratch (as per Raghavendra et al. who achieved 98.13% accuracy on a 1,426 image private test set [12]). The conversion of expert graders into deep models, as shown by Phene et al., have been found to be nearly as accurate on an adjudicated dataset (AUC 0.945 for referable

glaucomatous optic neuropathy) [13]. Transfer learning is competitive, as [14] combined active and transfer learning, and achieved an AUC of 0.9955 using ResNet on 8,433 images; Gómez-Valverde et al. compared several backbones on RIM-ONE with VGG19 achieving an AUC of 0.94 [15]; [16] used ResNet with screening and advanced disease being trained separately, and put on an AUC of 0.967 on 4,164 highly myopic cases and 14,575 non-highly-myopic cases; fine-tuned Inception-v3 to train for early and advanced disease with an AUC of 0.94 [17]; and [18] extended transfer learning to macular OCT thickness maps, outperforming classical machine learning baselines on 13,553 cases [18]. Demographic robustness has been addressed by Christopher et al., who's ResNet-50 reached an AUC of 0.91 across an ethnically diverse cohort of 14,822 images [19]. Serener and Serte compared GoogLeNet and ResNet-50 transfer learning for early and advanced glaucoma, with GoogLeNet achieving AUC 0.94 on advanced cases [20]. More recent segmentation-plus-classification pipelines have pushed reported metrics higher on ACRIMA specifically Sreng et al. combined DeepLabv3+ disc segmentation with a DenseNet classifier and reported 99.53% accuracy [21], while Elangovan and Nath reached 96.64% accuracy with a custom 18-layer CNN [22]. Efficient-backbone approaches such as Nawaz et al.'s EfficientDet-D0 pipeline have reported 97.2% accuracy on ORIGA while remaining deployable on commodity hardware [23]. Interpretability, typically via Grad-CAM [24], and clinician-facing deployment remain comparatively under-reported in this body of work.

Despite this rapid progress in the field, two gaps motivate the present study. First, few published systems are validated on South Asian cohorts despite documented ethnic and anatomical variation in optic-disc morphology that affects generalization. Second, many high-accuracy systems rely on computationally expensive ensembles that are difficult to host in low-bandwidth clinical environments. We therefore develop a region-adapted two-phase VGG16 pipeline that prioritises tractable training, Grad-CAM interpretability and a browser-based clinician dash-board for use in Pakistani ophthalmology services.

More recent work has shifted toward explainability, cross-modal fusion, and larger-scale meta-analysis. [25] provide a comprehensive review of deep learning and computer vision techniques for glaucoma detection, cataloguing the shift toward attention-based and transformer architectures. [26] proposed an automated diagnosis pipeline achieving strong performance on public fundus datasets while emphasizing computational efficiency for deployment. [27] demonstrated a ConvNeXt-based architecture with a convolutional block attention module (CBAM) for glaucoma detection in highly myopic populations, reporting AUC 0.894, and used Grad-CAM for interpretability in a manner directly comparable to the present study. [28] showed that a deep-learning image-quality filter substantially improves downstream automated primary-open-angle-glaucoma detection accuracy, highlighting preprocessing quality as an underexplored variable. An updated systematic review and meta-analysis, which combined, across 48 studies, the sensitivity of the fundus-photography-based deep learning glaucoma detection was 0.92 (95% CI: 0.89–0.94) and specificity was 0.93 (95% CI: 0.90–0.95) [29], providing a baseline by which single-center deep learning glaucoma detection studies can be compared. Jin et al. provide a lucid, comprehensive review of AI applications across glaucoma screening and diagnosis, reflecting on current limitations and outlining prospects for broader ophthalmic deployment [30]. Most recently, [31] reviewed computer-aided diagnosis systems for glaucoma and identified safety, reliability, interpretability, and explainability as the primary open gaps constraining clinical translation limitations that directly motivate the region-adapted, interpretability-focused design of the present study.

Research Novelty:

While the individual components used in this study VGG16 transfer learning, staged fine-tuning, and Grad-CAM interpretability are each individually well established in the

glaucoma-screening literature, their combination and framing in the present work differs from prior published systems in four respects.

First, unlike most transfer-learning studies that either freeze the backbone entirely or fine-tune it in a single stage, this work isolates the marginal contribution of a staged unfreezing schedule through a dedicated ablation (Table 1), quantifying a 7.15-percentage-point sensitivity gain and a corresponding halving of the false-negative rate attributable specifically to unfreezing the last four convolutional layers at a reduced learning rate. Few of the comparator studies in Table 2 report an equivalent within-study ablation isolating the effect of fine-tuning depth.

Second, the present system integrates model development with a deployable, clinician-facing web dashboard (FastAPI inference service and Next.js frontend) within the same study, rather than reporting model performance in isolation. Most of the high-accuracy systems reviewed in Section 1.1 report classification metrics without an accompanying deployment architecture suitable for non-specialist clinical use.

Third, the architectural choices in this study a single non-ensemble backbone, no separate segmentation stage, and a lightweight custom classification head were deliberately made to prioritise deployability on low-bandwidth, low-resource clinical infrastructure over incremental accuracy gains, in contrast to segmentation-plus-classification pipelines and multi-branch ensembles that achieve marginally higher accuracy at substantially higher computational and engineering cost.

Fourth, this study is explicitly positioned within a South Asian, and specifically Pakistani, clinical context. A parallel pilot cohort from the Sindh Institute of Ophthalmology and Visual Sciences (SIOVS), Hyderabad, is reserved for prospective regional validation, addressing the documented scarcity of glaucoma deep-learning systems evaluated on South Asian populations, where optic-disc morphology and imaging conditions may differ from the predominantly Western and East Asian cohorts underlying most published models.

Objectives:

Based on the gaps, specifically, measurable objectives of this research are:

To design a two-phase pipeline of transfer learning using different VGG layers for accuracy of at least 85% obtained on a balanced held-out test set ACRIMA, following some screening level thresholds set in the literature on deep-learning algorithms for glaucoma.

To measure, in the context of a controlled ablation, the isolated effect of Phase-2 fine-tuning (unfreezing the last four convolutional layers) by comparing to the baseline basing frozen-backbone, using accuracy, sensitivity, specificity, precision, and AUC-ROC.

To create the Grad-CAM interpretability maps for each test-set prediction and see if they were anatomically consistent with the optic-disc and peripapillary region which ophthalmologists use routinely in their practice.

To deploy the trained model behind a FastAPI inference service and a Next.js clinical dashboard that returns a glaucoma probability and Grad-CAM overlay to a non-specialist user in < 0.5-1.5 seconds per image, without the need for local GPU infrastructure. To create a reproducible open pipeline, suitable for prospective validation in larger clinical sample of Pakistan (SIOVS, Hyderabad), due to the lack of deep-learning systems for glaucoma validated in Asia.

Proposed Methodology:

This section provides an end-to-end pipeline for automated glaucoma classification which involves data acquisition, clinical deployment. The pipeline involves retinal fundus image preprocessing and augmentation, a two-step transfer-learning strategy using an inherent trunk network of VGG16, and finally a web-based clinical dashboard which can be interpretable by Grad-CAM for three years after training. The pipeline is summarized in figure 2 and will be described in detail in the following sections.

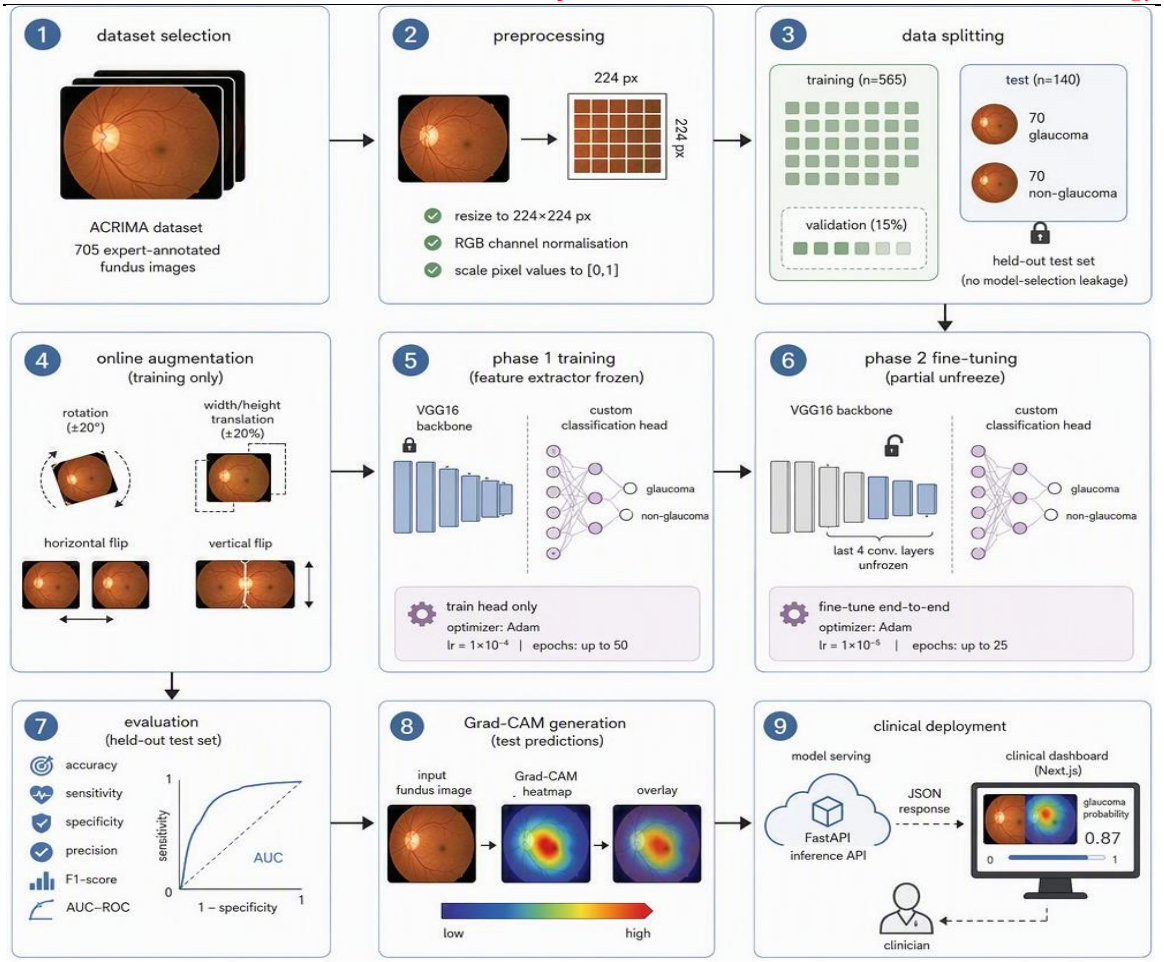

Figure 2. End-to-end methodology. Public fundus images are preprocessed and augmented, fed through a VGG16 backbone pre-trained on ImageNet, and classified by a custom dense head. A FastAPI inference service exposes the trained model to a Next.js clinical dashboard; Grad-CAM overlays support clinician interpretation.

Data Collection:

The ACRIMA dataset consists of 705 eye level images, each individually labelled, comprising 396 glaucomatous and 309 normal fundus images; no patient-level identifiers are provided that will allow multiple images for a patient to be identified as belonging to the same person. Images were partitioned at the image level, using standard practice for this dataset, via stratified random sampling to assign the images to training (70%), internal validation (20%), and held-out test sets (10%) with the aim of maintaining the overall proportions of the classes. Several orthogonal folds of the held-out test set (n = 140) were balanced by design (70 glaucomatous, 70 normal) while the internal validation split (15% of the 565-image training pool) used stratified sampling to obtain roughly equal numbers of glaucomatous (49) and normal (36) images in the validation set, and the balance (approximately 277 glaucomatous, 203 normal) in the training set. The numbers mentioned above are rounded up or down based on the stratified split routine and the time the rounds are split and are reported to the nearest image that will meet the 15% requirement and the pool's composition. The test partition was frozen prior to making model selection or hyper-parameter decisions.

Data Preprocessing:

All images were resized to 224×224 pixels in three channels RGB and normalized to be in range [0, 1] as required by the ImageNet pretrained backbone. The number of images was increased online by adding rotations of ±20◦, horizontal and vertical translations of ±20%,

in-place flips, and nearest-neighbor fill of any vacant areas. The validation and test images were not augmented, however, to ensure that the estimate of a model's performance was unbiased.

Model Architecture:

The proposed classification network consists of a pre-trained VGG16 network backbone and a lightweight binary glaucoma prediction head network. An example of this architecture can be seen in figure 3 where the five convolutional modules (VGG16) are connected to the classification blocks described below. The backbone used was VGG16 with ImageNet weights and without its fully connected top. A global average pooling layer was added followed by a dense layer of 512 rectified linear units with dropout rate of 50% and a dense layer of 256 rectified linear units with dropout rate of 30% and a sigmoid neuron with a single output providing the glaucoma probability. This was a complete network with 15,108,929 parameters. The model was implemented using TensorFlow 2.x, Keras.

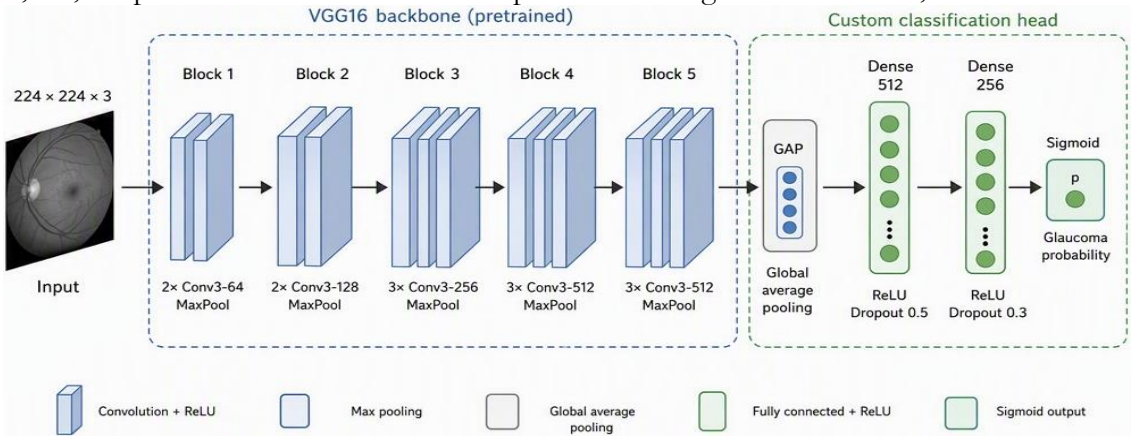

Figure 3. VGG16 backbone with the custom classification head used in this study. The five convolutional blocks of VGG16 are retained; the original fully connected top is replaced by global average pooling, two dense ReLU layers with dropout, and a single sigmoid output producing a glaucoma probability.

Two-Phase Training Strategy:

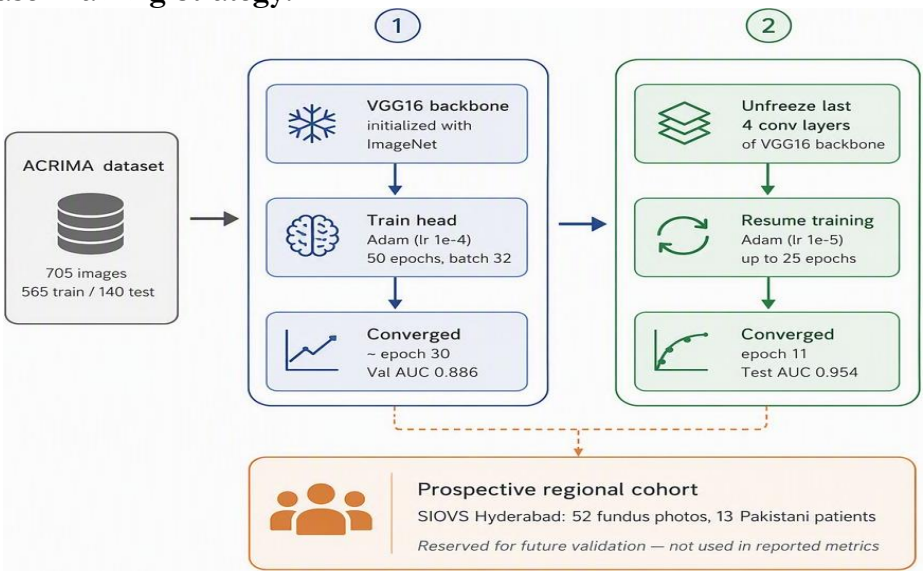

Figure 4. Two-phase transfer-learning schedule. Phase-1 trains the custom head with the VGG16 backbone fully frozen at lr 1 × 10−4. Phase-2 unfreezes the last four convolutional layers of VGG16 and continues training at lr 1 × 10−5 to specialize high-level filters to glaucoma-relevant features.

Training proceeded in two sequential phases to balance rapid convergence with careful specialization of pre-trained features. Figure 4 illustrates this schedule, showing the transition from a frozen-backbone phase to a fine-tuning phase with a reduced learning rate.

Phase-1 froze all convolutional layers of VGG16 and trained only the classification head, minimizing binary cross-entropy with Adam at a learning rate of $1 \times 10{-}4$, a batch size of 32, and a limit of 50 epochs. The objective of Phase-1 was to fit task-specific weights in the dense head while preserving the transferable visual features encoded by ImageNet pre-training.

Phase-2 unfroze the last four convolutional layers, which contain the highest-level filters in VGG16 and continued training at a reduced learning rate of $1 \times 10{-}5$ for up to 25 epochs. The deliberately smaller learning rate allowed these filters to specialize to glaucomatous features such as optic-disc cupping and neuroretinal-rim thinning without catastrophic forgetting. Both phases used

Early Stopping (patience 8 on validation loss, with best-weight restoration), ReduceLROnPlateau (factor 0.2, patience 4, minimum rate $1 \times 10{-}7$) and Model Checkpoint callbacks. Fine-tuning converged after 11 epochs.

Algorithm:

Algorithm 1: Two-Phase VGG16 Transfer Learning for Glaucoma Classification

Input: Dtrain, Dval, Dtest of fundus images (224×224×3) pre-trained VGG16 weights WImageNet
Output: fine-tuned classifier f ∗
1. f → VGG16(weights=WImageNet, include_top=False)
2. head → GAP → Dense(512, ReLU) → Dropout(0.5) → Dense(256, ReLU) → Dropout(0.3) → Sigmoid
3. f → Sequential(f , head)
4. Augmentation: rotate ±20◦, shift ±20%, flip H/V
5. # Phase-1: train head with backbone frozen
6. freeze(f .base_layers)
7. train f with Adam(lr=10−4), BCE loss, up to 50 epochs, batch_size=32, class_weight = inverse_frequency(Dtrain) callbacks = {EarlyStopping(p=8), ReduceLROnPlateau, ModelCheckpoint}
8. # Phase-2: fine-tune last 4 convolutional layers
9. unfreeze(f .base_layers[-4:])
10. resume training with Adam(lr=10−5) for up to 25 epochs
11. evaluate f on Dtest → Accuracy, Sensitivity, Specificity, Precision, F1, AUC-ROC
12. generate GradCAM(f , x) for x ∈ Dtest
13. return f ∗

Evaluation Metrics:

Performance on the 140-image held-out test set was reported with accuracy, precision, sensitivity, specificity, F1-score and the area under the receiver operating characteristic curve (AUC-ROC). The pre-specified acceptance criterion was sensitivity ≥ 85%, consistent with screening-level reporting thresholds in the glaucoma literature.

95% confidence intervals for proportion-based metrics (accuracy, sensitivity, specificity, and precision) were computed using the Wilson score interval. The confidence interval for AUC-ROC was estimated using the Hanley–McNeil method, based on the reported AUC and the number of positive and negative test cases ($n1 = n2 = 70$), as raw prediction scores were not retained for post-hoc DeLong analysis.

Clinical Deployment:

The trained model was packaged behind a FastAPI inference service deployed on Railway, and a Next.js/React web dashboard was deployed on Vercel. A clinician uploads a fundus image through the browser, receives a glaucoma probability and a Grad-CAM overlay within 0.5 to 1.5 seconds, and can download a report. Grad-CAM heatmaps were generated from the final convolutional block and superimposed on the original image in the JET colour scheme.

Results:

Overall Classification Performance:

The fine-tuned model's diagnostic performance was evaluated on the 140-image held-out test set using standard classification metrics. Figure 5 summarizes these results, presenting the confusion matrix alongside the ROC curve. On the 140-image held-out test set, the fine-tuned network achieved an accuracy of 88.57% (95% CI: 82.2–92.8%), sensitivity 92.86% (95% CI: 84.3–96.9%), specificity 84.29% (95% CI: 74.0–91.0%), precision 85.53% (95% CI: 75.9–91.7%), F1-score 89.04%, and AUC-ROC 0.954 (95% CI: 0.918–0.990, Hanley–McNeil approximation). The confusion matrix showed 65 true positives, 59 true negatives, 11 false positives and 5 false negatives, corresponding to a false-negative rate of 7.1% and a false-positive rate of 15.7%. The high AUC-ROC, which is threshold-independent, indicates strong discrimination and permits re-calibration of the decision point if a different sensitivity/specificity balance is preferred in a particular screening setting.

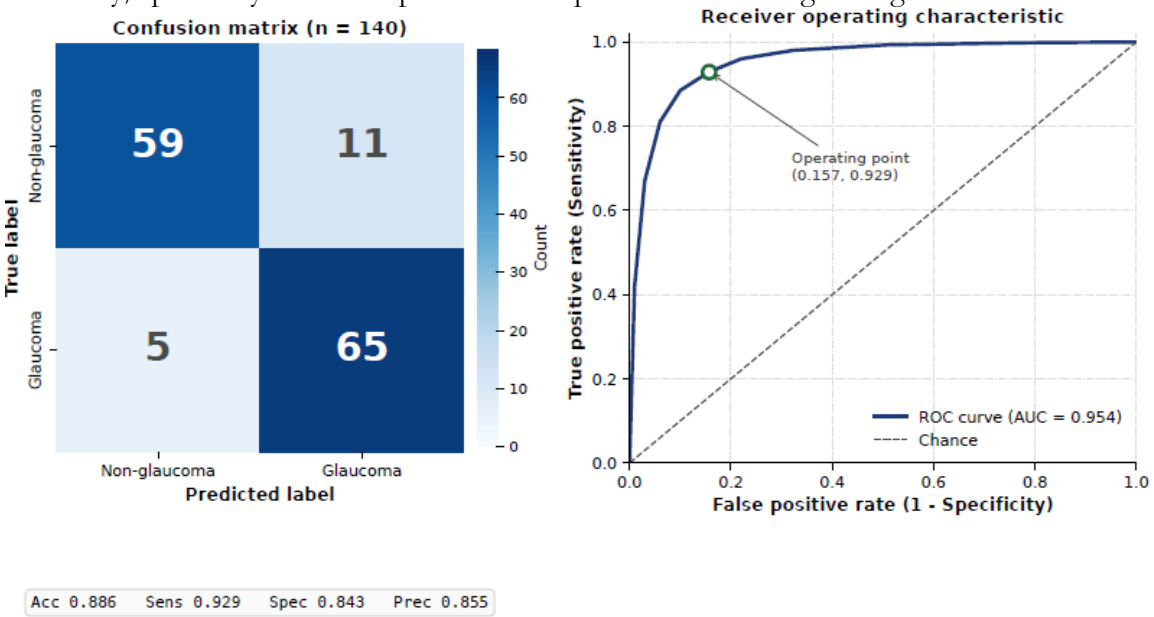

Figure 5. Diagnostic performance of the fine-tuned VGG16 model on the 140-image held-out test set. Left: confusion matrix (65 TP, 59 TN, 11 FP, 5 FN). Right: receiver operating characteristic curve with AUC = 0.954.

Training Dynamics:

Model convergence was monitored across both training phases using loss and accuracy on the training and validation sets. Figure 6 presents these trajectories, marking the transition between Phase-1 and Phase-2.

The Phase-1 head converged within roughly 30 epochs with validation accuracy stabilizing near 88%. Phase-2 fine-tuning produced a smooth additional improvement through epoch 10, with EarlyStopping halting training at epoch 11, consistent with the small-learning-rate regime used to avoid disturbing the pre-trained lower layers.

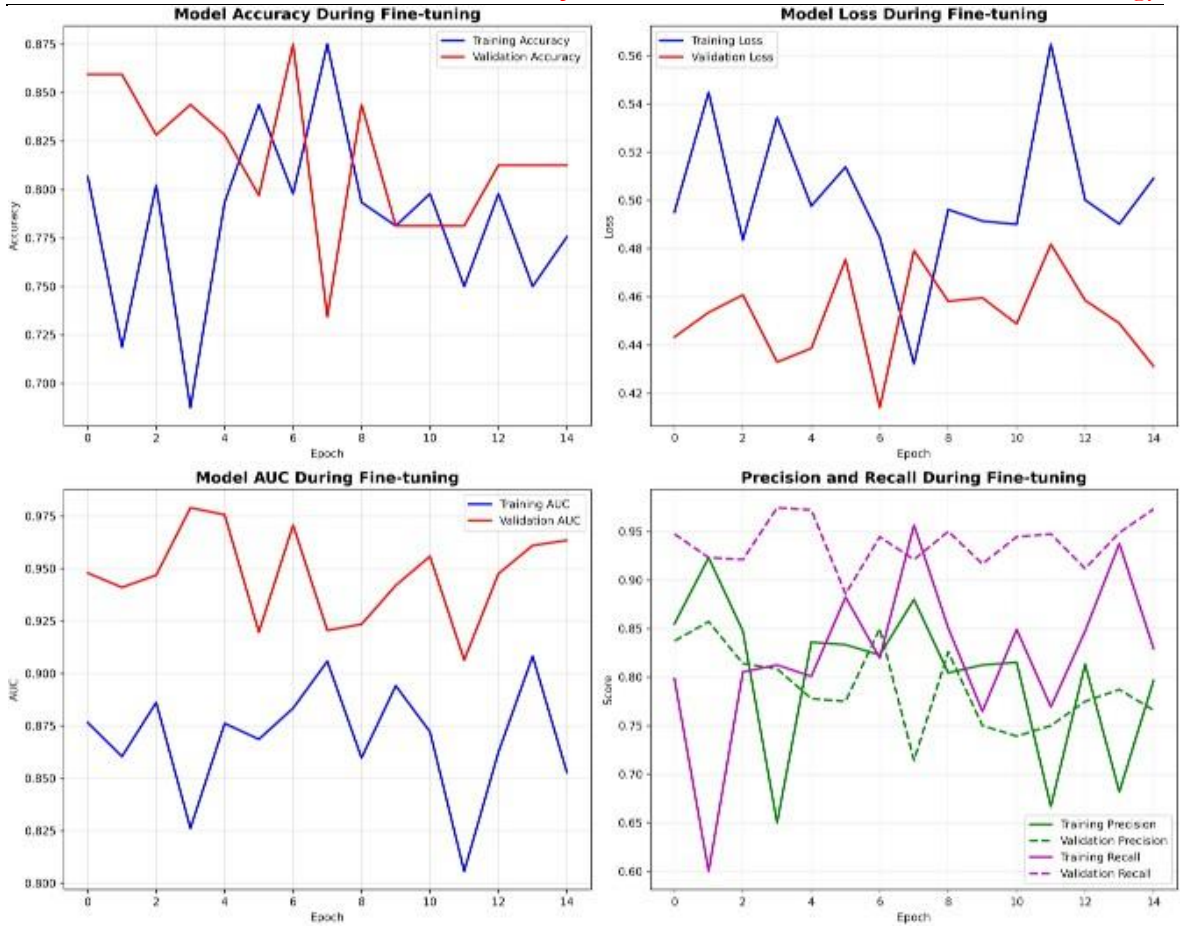

Figure 6. Training and validation loss and accuracy trajectories across Phase-1 (frozen backbone, lr 1 × 10−4) and Phase-2 (last four convolutional layers unfrozen, lr 1 × 10−5).

Ablation Study:

The isolated contribution of Phase-2 fine-tuning is reported in Table 1 and visualized in Figure 7. Unfreezing the last four convolutional layers and resuming training at 1 × 10−5 raised sensitivity by 7.15 percentage points while leaving overall accuracy unchanged. The simultaneous drop in precision from 90.91% to 85.53% is the expected screening trade-off the fine-tuned model is more willing to flag borderline cases, which lowers the false-negative rate at the cost of a higher false-positive rate. Additional configurations were also compared during development: training the same head without augmentation converged to a validation accuracy approximately 4–6 percentage points lower, and unfreezing the entire VGG16 backbone from the start destabilized training and produced worse validation metrics, confirming the value of both the augmentation pipeline and the staged unfreezing schedule.

Table 1. Ablation of the two-phase training schedule on the held-out test set.

Configuration	Accuracy	Sensitivity	Specificity	Precision	AUC-ROC
Phase-1 only (head, backbone frozen)	88.57%	85.71%	91.43%	90.91%	0.951
Phase-2 (last 4 conv layers unfrozen)	88.57%	92.86%	84.29%	85.53%	0.954
Phase-1 without augmentation (dev)	~84%	~82%	~86%	~85%	0.92
Full-backbone unfreeze from start (dev)	~79%	~75%	~83%	~80%	0.87

Note: Rows 3–4 report approximate development-time metrics; exact values with confidence intervals were not retained as these configurations were exploratory checks rather than final reported models. Precise re-evaluation on the frozen test set is left for future work.

A formal paired significance test (e.g., McNemar's test) between Phase-1 and Phase-2 predictions was not performed, as per-image Phase-1 predictions were not retained after Phase-2 superseded it (see Limitation 8). Although the sensitivity confidence intervals for Phase-1 (75.7–92.0%) and Phase-2 (84.3–96.9%) overlap, the 7.15-percentage-point point-estimate gain and the corresponding halving of the false-negative rate (Table 1) are suggestive of a meaningful rather than incidental improvement. We recommend that future iterations of this pipeline retain raw prediction outputs from every training phase to enable rigorous paired significance testing.

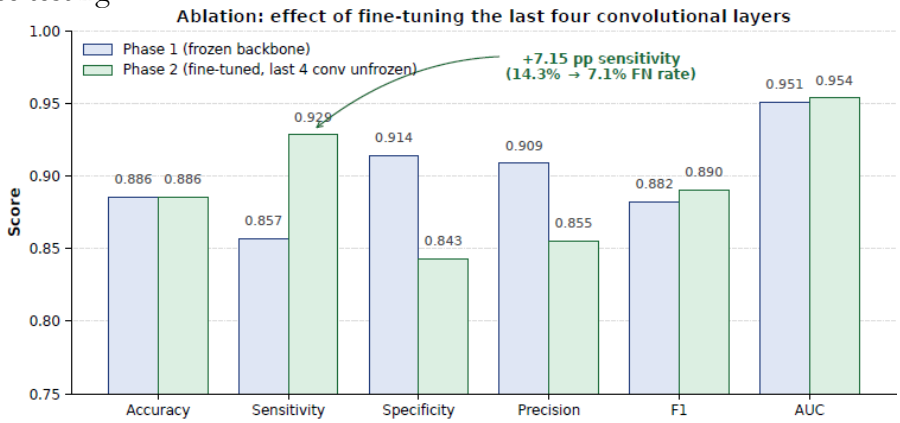

Figure 7. Metric-by-metric comparison between the Phase-1 frozen-backbone model and the Phase-2 fine-tuned model, illustrating the sensitivity gain produced by unfreezing the last four convolutional layers.

Comparison with Published Work:

Table 2 situates our result against eight representative published deep-learning glaucoma classifiers. The present system is directly comparable to the ACRIMA-evaluated rows: and Elangovan and Nath report higher accuracy figures but rely on segmentation-plus-classification pipelines and substantially longer training regimes. Transfer-learning baselines at comparable scale notably[8] multi-dataset Xception[15] VGG19 on RIM-ONE and [14] ResNet with active learning on 8,433 images report sensitivities in the 87–99% range, within which our 92.86% falls.

Table 2. Comparison with representative published glaucoma-detection systems

Study (Year)	Dataset	Architecture	Sens.	Spec.	AUC
[7]	Private (n=48,116)	Inception-v3	95.6%	92.0%	0.986
[8]	5 public (incl. ACRIMA)	Xception (TL)	93.5%	85.8%	0.961
[12]	Private (n=1,426)	18-layer CNN (scratch)	98.0%	98.3%	—
[14]	Private (n=8,433)	ResNet + active/TL	99.2%	93.0%	0.996
[15]	RIM-ONE	VGG19 (TL)	87.0%	89.0%	0.94
[11]	ORIGA	RCNN + CNN	—	—	0.874
[21]	ACRIMA	DeepLabv3+ + DenseNet	—	—	0.9998
[22]	ACRIMA	Custom 18-layer CNN	96.1%	97.4%	—
Proposed study (Phase-2)	ACRIMA (n=140 test)	VGG16 (TL) + 2-phase FT	92.86%	84.29%	0.954

Interpretability via Grad-CAM:

Grad-CAM heatmaps generated from the final convolutional layer were visually inspected across the test set and qualitatively localized to the optic-disc and peripapillary region in the large majority of cases, in apparent spatial correspondence with the cup-to-disc ratio and neuroretinal-rim features ophthalmologists use in routine practice. This assessment was qualitative; a systematic quantitative validation (e.g., overlap with expert-annotated disc/cup masks, or blinded clinician concordance rating) was not performed in the present study and is identified below as a priority for future work.

Discussion:

This work delivers a two-phase transfer-learning pipeline whose sensitivity of 92.86% clears the screening-utility threshold commonly cited in the glaucoma deep-learning literature, with an AUC-ROC of 0.954 on a balanced held-out ACRIMA test set. The principal methodological contribution is the staged frozen-then-fine-tuned schedule: Phase-1 stabilized task-specific weights in the dense head without disturbing transferable Image-Net features, and Phase-2 specialized only the highest-level filters at a learning rate one order of magnitude lower, avoiding catastrophic forgetting. Beyond the pre-specified 85% acceptance threshold, the 92.86% sensitivity achieved here compares favorably with the pooled sensitivity of 0.92 (95% CI: 0.89–0.94) reported across 48 studies in the most recent systematic review and meta-analysis of fundus-photography-based deep learning glaucoma detection, and sits within the 87–99% range reported by comparable single-backbone transfer-learning systems evaluated on public datasets. Screening programs conventionally tolerate a higher false-positive rate in exchange for a low false-negative rate, since missed glaucoma cases risk irreversible vision loss before diagnosis, whereas false positives are resolved by routine specialist referral; our sensitivity/specificity trade-off (92.86%/84.29%) reflects this screening-oriented design philosophy rather than a diagnostic-replacement one. The ablation (Table 1) isolates this contribution: fine-tuning converted a 14.3% false-negative rate into a 7.1% false-negative rate, which is the failure mode that matters most for a screening tool.

The achieved metrics are competitive with transfer-learning baselines of comparable scale on public datasets and sit below specialized segmentation-plus-classification pipelines that trade computational cost for marginal accuracy gains. The design choices of this study a single backbone, no ensemble, no separate segmentation stage, full open-source reproducibility prioritize deployability in low-resource environments over leaderboard position.

Clinical Dashboard and Translational Utility:

The trained model was deployed as an accessible clinical decision support dashboard via the internet, moving beyond the 'notebook' prototype. The resulting clinician facing UI is displayed in the three main views of figure 8. The architecture separates a FastAPI inference service from a NextJS application running as a front-end, eliminating the need for clinical sites to have GPU infrastructure setup on-site. The final result is within seconds after a non-specialist user uploads a fundus image, with a glaucoma probability being shown and a Grad-CAM overlay as well, which can be exported to the patient record. This reduces the barrier of adoption, especially in places like SIOVS Hyderabad where the cost of the client's hardware device is a critical constraint. Figure 8 illustrates the landing page, upload workflow and result view.

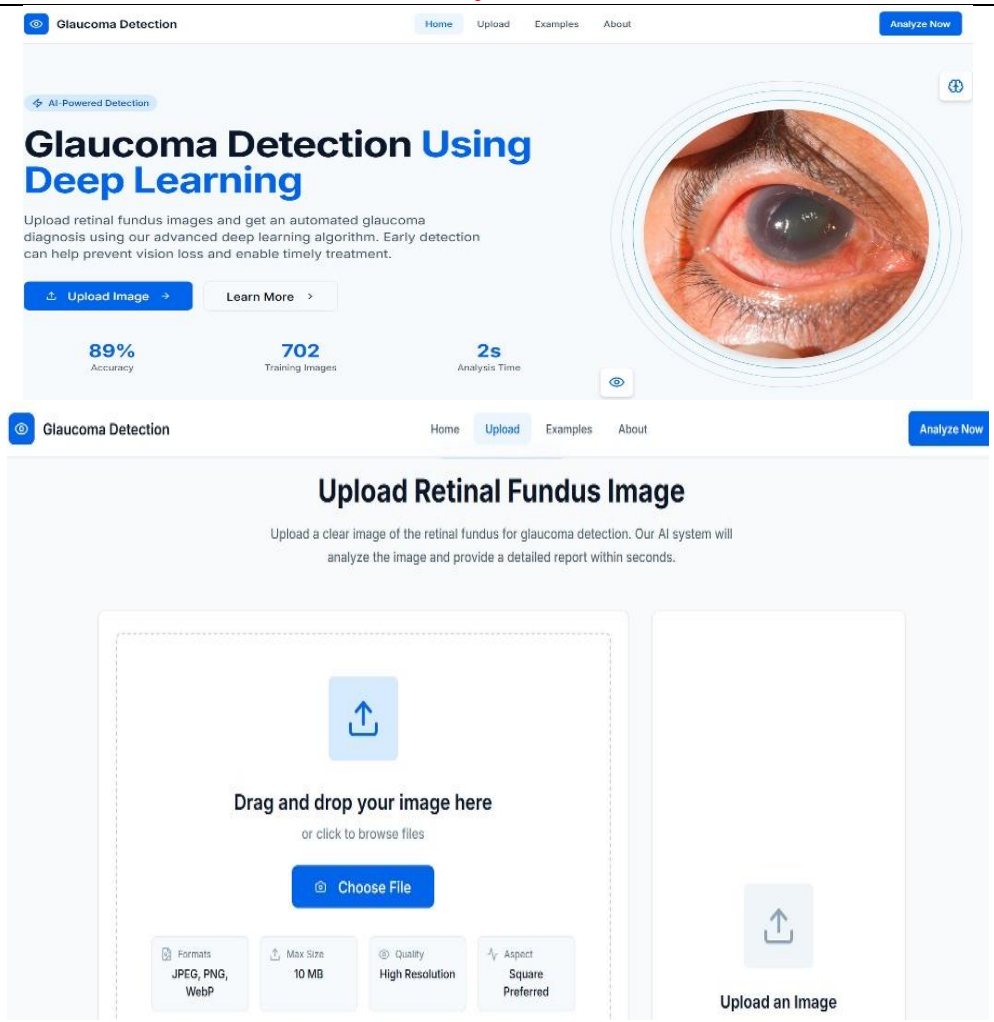

Figure 8. Clinician-facing dashboard. Top: landing page. Middle: fundus-image upload view. Bottom: result view showing predicted glaucoma probability and Grad-CAM overlay highlighting the optic-disc region that drove the prediction.

Limitations and Future Work:

Several limitations constrain the conclusions that can be drawn. First, quantitative metrics were obtained on ACRIMA only; the SIOVS Hyderabad pilot cohort is framed here as a prospective regional adaptation target and has not yet contributed to the reported numbers. A regionally trained model evaluated on SIOVS data will be required before any specific performance claim on Pakistani patients is defensible. Second, the system performs binary classification only and does not stage disease severity, which limits its use in treatment-planning. Third, performance on subtle early-stage changes was not characterized separately. Fourth, the cloud deployment requires internet connectivity; an offline, on-device variant based on model compression is an immediate next step. Fifth, formal multi-reader comparison against ophthalmologist consensus and regulatory review are prerequisites for any clinical deployment. Sixth, the confidence intervals for AUC-ROC were recalculated using the Hanley–McNeil approximation and not DeLong's method, because in future work there is a possibility to retain raw prediction probabilities for held-out test set so that the confidence intervals can be calculated as non-parametric, or by the boostrapping approach for the F1 score, respectively. Seventh, patient-level and eye-level identifiers are not published by ACRIMA, and thus, there is no information to rule out that the two images were taken in the same eye on two separate occasions; this explains why splitting was done at the image level, as

was done in all previous ACRIMA-based studies. SIOVS Hyderabad cohort would not have this constraint as patient data identifiers will be available and strict patient-level split validation would be done on a prospective patient basis. It was not possible to make a formal statistical test (such as a McNemar's test) between predictions from Phase-1 and predictions from Phase-2 because the predictions for each image were discarded after Phase-2 training took over, though in future the raw predictions from each of the various training stages will be saved for such tests to be carried out. The values reported in the augmentation-ablation and full-backbone-unfreeze configurations in Table 1 are approximate development-time experiments values, not fully re-evaluated models with a confidence interval; future "replicates" should capture all of the metrics for all of the ablation arms from the beginning. Tenth, though, the interpretability of Grad-CAM was not evaluated quantitatively; it was judged qualitative, by visual inspection – meaning, either by means of a pooling of optic-disc and optic-cup segmentation masks developed by experts, or by comparison between a blind clinician and user on a defined set of images. Formal quantitative validation of the fidelity of the saliency map, by an ophthalmologist who is not part of the SIOVS study, is planned in the SIOVS-Hyderabad prospective validation study.

Implications and Recommendations:

Clinical implications: The high sensitivity (92.86%) suggests the use of the system as a first-line ancillary triage tool for specialist referral and not a replacement for first-line diagnostic systems, and the 7.1% false-negative rate suggests periodic re-examination of high-risk patients in the clinical area. The use of Grad-CAM overlays assists general staff to gain confidence and understand referrals.

Regional implications: All metrics are based on ACRIMA: The system is not a regionally tested one and only the 'reserved' Hyderabad cohort of SIOVS is the planned route to regional validation in the time being.

Recommendations:

Validating Grad-CAM localization quantities, instead of relying on qualitative visual inspection alone, by comparing against optic disc/cup masks, or blinded clinician concordance ratings.

Review the responses raised in the patient-level splitting mode for validation of SIOVS Hyderabad.

Store the raw predictions per image throughout training to use for paired significance testing (such as McNemar's test) and DeLong AUC confidence intervals.

Discuss model compression for Offline/On Device Deployment in Low Connectivity Clinics.

Link the severity staging of the classifier to the planning for treatment.

Conduct several multiple-reader concordance studies in preparation for clinical introduction, and regulatory review.

Conclusion:

A two-phase transfer-learning system on VGG16 had a sensitivity of 92.86%, accuracy of 88.57 %, and AUC-ROC of 0.954 for binary glaucoma classification on retinal fundus images on the ACRIMA test set along with the interpretability using Grad-CAM and a clinician facing web dashboard. The pipeline is easy to compute, can be reproduced from the data and roll out to any commodity cloud infrastructure, which offers a useful platform for region-adapted glaucoma screening in Pakistani population until prospective testing on SIOVS Hyderabad population could confirm or refute its validity.

CRediT Authorship Contribution Statement:

Shafiya Qadeer Memon: Conceptualization, Supervision, Writing original draft (Corresponding author).

Rafay Shakeel: Methodology, Software.

Sachal Sahito: Software, Validation.
Amirita Dewani: Software, Validation.
Areeba Syed: Data curation, Validation and editing.
Mehran Memon: Clinical supervision, Resources, Writing – review & editing.

Declaration of Competing Interest:

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

Funding:

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability:

The ACRIMA dataset used in this study is publicly available from its original publication. Source code is available from the corresponding author on reasonable request.

Acknowledgements:

The authors thank the Department of Software Engineering, Mehran University of Engineering and Technology, Jamshoro, and the Sindh Institute of Ophthalmology and Visual Sciences (SIOVS), Hyderabad, for institutional and clinical support.

References:

[1] Y. C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C. Y. Cheng, "Global prevalence of glaucoma and projections of glaucoma burden through 2040: A systematic review and meta-analysis," *Ophthalmology*, vol. 121, no. 11, pp. 2081–2090, Nov. 2014, doi: 10.1016/j.ophtha.2014.05.013.

[2] M. S. M. Zahid Hussain Awan, P.S. Mahar, "Blindness and Poverty," *Pak J Ophthalmol*, vol. 27, no. 3, pp. 165–170, Nov. 2011, Accessed: Aug. 25, 2026. [Online]. Available: https://pjo.org.pk/index.php/pjo/article/view/489/418

[3] "WHO Releases World Report on Vision." Accessed: Aug. 25, 2026. [Online]. Available: https://cureblindness.org/news/who-releases-world-report-on-vision?utm_source=google&utm_medium=cpc&utm_campaign=dynamic-grant&utm_id=10038769322&utm_term=&utm_content=730060842429&utm_source_platform=google-ads&utm_creative_format=text&utm_marketing_tactic=acquisition&c_src=WEB0126GG&gad_source=1&gad_campaignid=10038769322&gbraid=0AAAADr1f9Wi1U4QYqJFcN-5q2tIIgHML&gclid=CjwKCAjw-rTUBhAiEiwADv8gBD60c7fknw2kVB0Qx0-cZZW1HEy3il1nXcs9NHu4_0-5ZrAEUilMqxoClT0QAvD_BwE

[4] R. N. Weinreb, T. Aung, and F. A. Medeiros, "The pathophysiology and treatment of glaucoma: a review," *JAMA*, vol. 311, no. 18, pp. 1901–1911, May 2014, doi: 10.1001/JAMA.2014.3192.

[5] "Glaucoma: What It Is, Symptoms, Causes, Types & Treatment." Accessed: Aug. 25, 2026. [Online]. Available: https://my.clevelandclinic.org/health/diseases/4212-glaucoma

[6] D. S. W. Ting *et al.*, "Artificial intelligence and deep learning in ophthalmology," *Br. J. Ophthalmol.*, vol. 103, no. 2, pp. 167–175, Feb. 2019, doi: 10.1136/BJOPHTHALMOL-2018-313173.

[7] Z. Li, Y. He, S. Keel, W. Meng, R. T. Chang, and M. He, "Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs," *Ophthalmology*, vol. 125, no. 8, pp. 1199–1206, Aug. 2018, doi: 10.1016/j.ophtha.2018.01.023.

[8] A. Diaz-Pinto, S. Morales, V. Naranjo, T. Köhler, J. M. Mossi, and A. Navea, "CNNs for automatic glaucoma assessment using fundus images: an extensive validation,"

Biomed. Eng. OnLine 2019 181, vol. 18, no. 1, pp. 29-, Mar. 2019, doi: 10.1186/S12938-019-0649-Y.

[9] J. I. Orlando *et al.*, "REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Med. Image Anal.*, vol. 59, p. 101570, Jan. 2020, doi: 10.1016/J.MEDIA.2019.101570.

[10] H. Fu *et al.*, "Disc-Aware Ensemble Network for Glaucoma Screening from Fundus Image," *IEEE Trans. Med. Imaging*, vol. 37, no. 11, pp. 2493–2501, Nov. 2018, doi: 10.1109/TMI.2018.2837012.

[11] M. N. Bajwa *et al.*, "Two-stage framework for optic disc localization and glaucoma classification in retinal fundus images using deep learning," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, Jul. 2019, doi: 10.1186/S12911-019-0842-8.

[12] U. Raghavendra, H. Fujita, S. V. Bhandary, A. Gudigar, J. H. Tan, and U. R. Acharya, "Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images," *Inf. Sci. (Ny).*, vol. 441, pp. 41–49, May 2018, doi: 10.1016/J.INS.2018.01.051.

[13] S. Phene *et al.*, "Deep Learning and Glaucoma Specialists: The Relative Importance of Optic Disc Features to Predict Glaucoma Referral in Fundus Photographs," *Ophthalmology*, vol. 126, no. 12, pp. 1627–1639, Dec. 2019, doi: 10.1016/j.ophtha.2019.07.024.

[14] R. Hemelings *et al.*, "Accurate prediction of glaucoma from colour fundus images with a convolutional neural network that relies on active and transfer learning," *Acta Ophthalmol.*, vol. 98, no. 1, pp. e94–e100, Feb. 2020, doi: 10.1111/AOS.14193.

[15] J. J. Gómez-Valverde *et al.*, "Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning," *Biomed. Opt. Express*, vol. 10, no. 2, p. 892, Feb. 2019, doi: 10.1364/BOE.10.000892.

[16] N. Shibata *et al.*, "Development of a deep residual learning algorithm to screen for glaucoma from fundus photography," *Sci. Reports 2018 81*, vol. 8, no. 1, pp. 14665-, Oct. 2018, doi: 10.1038/s41598-018-33013-w.

[17] J. M. Ahn, S. Kim, K. S. Ahn, S. H. Cho, K. B. Lee, and U. S. Kim, "A deep learning model for the detection of both advanced and early glaucoma using fundus photography," *PLoS One*, vol. 13, no. 11, p. e0207982, Nov. 2018, doi: 10.1371/JOURNAL.PONE.0207982.

[18] R. Asaoka *et al.*, "Using Deep Learning and Transfer Learning to Accurately Diagnose Early-Onset Glaucoma From Macular Optical Coherence Tomography Images," *Am. J. Ophthalmol.*, vol. 198, pp. 136–145, Feb. 2019, doi: 10.1016/j.ajo.2018.10.007.

[19] M. Christopher *et al.*, "Performance of Deep Learning Architectures and Transfer Learning for Detecting Glaucomatous Optic Neuropathy in Fundus Photographs," *Sci. Reports 2018 81*, vol. 8, no. 1, pp. 16685-, Nov. 2018, doi: 10.1038/s41598-018-35044-9.

[20] A. Serener and S. Serte, "Transfer learning for early and advanced glaucoma detection with convolutional neural networks," *TIPTEKNO 2019 - Tip Teknol. Kongresi*, Oct. 2019, doi: 10.1109/TIPTEKNO.2019.8894965.

[21] S. Sreng, N. Maneerat, K. Hamamoto, and K. Y. Win, "Deep Learning for Optic Disc Segmentation and Glaucoma Diagnosis on Retinal Images," *Appl. Sci. 2020, Vol. 10, Page 4916*, vol. 10, no. 14, p. 4916, Jul. 2020, doi: 10.3390/APP10144916.

[22] P. Elangovan and M. K. Nath, "Glaucoma assessment from color fundus images using convolutional neural network," *Int. J. Imaging Syst. Technol.*, vol. 31, no. 2, pp. 955–971, Jun. 2021, doi: 10.1002/IMA.22494;CTYPE:STRING:JOURNAL.

[23] M. Nawaz *et al.*, "An Efficient Deep Learning Approach to Automatic Glaucoma Detection Using Optic Disc and Optic Cup Localization," *Sensors 2022, Vol. 22, Page*

434, vol. 22, no. 2, p. 434, Jan. 2022, doi: 10.3390/S22020434.

[24] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *Int. J. Comput. Vis.*, 2016, [Online]. Available: https://arxiv.org/abs/1610.02391

[25] Karen Simonyan, Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556*, 2015, [Online]. Available: https://arxiv.org/abs/1409.1556

[26] A. Shoukat, S. Akbar, S. A. Hassan, S. Iqbal, A. Mehmood, and Q. M. Ilyas, "Automatic Diagnosis of Glaucoma from Retinal Images Using Deep Learning Approach," *Diagnostics (Basel, Switzerland)*, vol. 13, no. 10, May 2023, doi: 10.3390/DIAGNOSTICS13101738.

[27] Y. Y. Chiang, C. L. Chen, and Y. H. Chen, "Deep Learning Evaluation of Glaucoma Detection Using Fundus Photographs in Highly Myopic Populations," *Biomedicines*, vol. 12, no. 7, p. 1394, Jul. 2024, doi: 10.3390/BIOMEDICINES12071394/S1.

[28] B. Chuter *et al.*, "Deep Learning Identifies High-Quality Fundus Photographs and Increases Accuracy in Automated Primary Open Angle Glaucoma Detection," *Transl. Vis. Sci. Technol.*, vol. 13, no. 1, Jan. 2024, doi: 10.1167/TVST.13.1.23.

[29] X. C. Ling *et al.*, "Deep Learning in Glaucoma Detection and Progression Prediction: A Systematic Review and Meta-Analysis," *Biomedicines*, vol. 13, no. 2, p. 420, Feb. 2025, doi: 10.3390/BIOMEDICINES13020420/S1.

[30] Y. Jin, L. Liang, J. Li, K. Xu, W. Zhou, and Y. Li, "Artificial intelligence and glaucoma: a lucid and comprehensive review," *Front. Med.*, vol. 11, 2024, doi: 10.3389/FMED.2024.1423813.

[31] Y. Hagiwara, O.-A. Ciora, M. Monnet, G. Lancho, and J. M. Lorenz, "AI-Driven Approaches for Glaucoma Detection -- A Comprehensive Review," Oct. 2024, Accessed: Aug. 25, 2026. [Online]. Available: https://arxiv.org/pdf/2410.15947

Copyright © by authors and 50Sea. This work is licensed under Creative Commons Attribution 4.0 International License.