A Reproducible Deep Learning Framework for Explainable Cataract Screening from Retinal Fundus Images

M. Z. Aziz [1], A. Zahid [2]

[1] *Department of Computing and Emerging Technologies, Emerson University Multan, Multan, Pakistan*
[2] *Beaconhouse College Program (BCP), Multan, Pakistan*

[1]zahid.aziz@eum.edu.pk

Abstract- Cataract remains a major cause of preventable low vision, particularly in communities with limited ophthalmic services. In this study, a research prototype of end-to-end Deep Learning model for binary classification of cataract from retinal fundus images was developed and tested. This contribution is not a new backbone architecture, but rather a reproducible pipeline, which combines the data integrity checking, ophthalmic images standardization, class-aware transfer learning, validation-only threshold selection, statistical evaluation, explainable artificial intelligence, and deployable single image inference. The public collection on Kaggle was examined and the duplicate repository tree was removed and 400 unique images (300 normal, 100 cataract) were isolated. Each image was then resized to 224×224 RGB with black padding in order to preserve aspect-ratio and CLAHE was used to enhance the images. The training, validation and test images were acquired by a fixed stratified split, which resulted in 280 training, 60 validation and 60 test images. The classes were trained using the same class weighted fashion in ResNet50, VGG19 and InceptionV3. Both VGG19 and ResNet50 model got 88.33% test accuracy at the traditional 0.50 threshold, with VGG19 having the highest ROC-AUC (0.9259) and specificity (95.56%) while ResNet50 had higher sensitivity (73.33%) and F1-score (75.86%). A validation-only Youden-J threshold of 0.03 resulted in a shift of the final operating point of VGG19 towards screening sensitivity 86.67%, with accuracy of 78.33%, specificity of 75.56%, F1-score of 66.67%, ROC-AUC of 0.9259, and PR-AUC of 0.8752. Two thousand iterations of the stratified bootstrap, paired comparisons, calibration analysis and review of the confusion matrix were performed. All the test images were explained using the grad-CAM, but the faithfulness using occlusion was mixed so it was interpreted with caution. The package reproduced exactly all the test probabilities and decisions and it is now possible to do quality-aware single-image screening. This system should not be considered to be a clinically validated system, but rather an explainable research prototype. It has a clear and strong advantage in the seamless integration of the following features: model comparison, threshold trade-offs, error analysis, deployment auditing, and all on a limited fundus dataset.

Keywords- Cataract Detection, Retinal Fundus Photograph, Transfer Learning, VGG19, ResNet50, InceptionV3, Grad-CAM, Threshold Optimization, Medical Image Classification.

I. INTRODUCTION

Cataract is a degenerative loss of transparency of the crystalline-lens, which causes less light to reach the retina both in terms of quantity and quality. Most affected eyes can be treated to restore sight, but the timing of treatment to make a difference depends on the ability to identify visuo-significant disease and transfer patients before they become disabled. Cataract remains a major cause of visual loss and distance vision impairment as listed by World Health Organization and they estimate tens of millions of people are visually impaired, and a significant proportion of these who will require cataract surgery cannot access it [1-2]. This burden is unevenly distributed across populations and healthcare settings. Rural populations, older people, women and in lower resource health systems, examination delays, travel and lack of trained ophthalmologists are more likely to occur. All these conditions render cataract screening an access as well as a diagnosis problem.

The conventional assessment is mainly performed by slit-lamp bio-microscopy, visual-acuity testing, ophthalmoscopy and lens-opacity grading done by trained clinicians. They are clinically relevant, require equipment, the expertise of the examiner and physical access. There is also a possibility of subjective grading in the process of grading,

especially in the vicinity of the classification boundaries. Digital retinal fundus photography cannot take the place of direct lens examination but cataract causes decreased contrast in the image, decreased vascular detail and diffuse blur. This indirect signal has prompted a study to see if fundus images taken for other screening programs can be used to opportunistically detect cataracts [3-4]. New hand-held slit lamp systems for smartphones and portables have increased the type of images that can be captured at non-tertiary health care centers [5], [10] which further strengthens the idea of using a scalable computer-assisted triage system.

Previous cataract detection methods involved hand-designed features such as texture features, wavelet, histogram and morphological features, and were then classified by support vector machines or other shallow classifiers. Such methods showed technical feasibility and were found to be very dependent on the acquisition conditions and would need to be manually selected. Transformed the field were deep CNN that learned hierarchical representation directly from pixels. A recent number of studies have shown good results with fundus photos, slit lamp images, optical coherence tomography (OCT) of the anterior segment and video frames [3-17]. But the accuracy reported alone can be deceptive due to the variations amongst the different studies in the image modality, the definition of classes, the number of patients, the partitioning of patients at the image level, the severity of the disease, and the external validation. A model that is trained on a small public fundus collection thus can't be equated with a multicenter system with tens of thousands of clinically graded images.

Transfer learning can be applied when there is limited amount of medical data. A network trained on ImageNet can thus leverage the generic features of edges, textures, shapes and compositions learned at the lower levels and can have higher level adjustments to the target problem. ResNet50 has the advantage of having residual connections which enable training of deep models, VGG19 has a uniform stack of small convolutions, while InceptionV3 is capable of capturing patterns at multiple receptive-field scales. These architectures are still valuable baselines to explore various trade-offs between the number of features, number of parameters and stability when tuning. Additionally, recent ophthalmic studies have focused on image-quality assessment, data augmentation, calibration, explainability and transparent reporting instead of a single, headline score [18-42].

This work started from a more general synopsis which suggested analyzing cataracts in the retinal fundus image and in the anterior-segment lens image. However, the experiment that is completed is smaller, less robust: binary classification of Normal versus Cataract is only valid on the retinal fundus dataset available. Extraction, cryptographic verification and audit was performed on the dataset obtained from Kaggle. Both trees of folders had the same 601 images; this was verified by comparing the two folder trees using SHA-256, and the duplicate one was removed. 300 normal and 100 cataract images were considered as the remaining 101 glaucoma and 100 retinal-disease images were removed. This is because results from an experiment performed in the fundus only, cannot be reported for anterior-segment classification, multimodal fusion or severity grading.

The proposed experimental pipeline is based on the following experimental steps: Resizing all images to a fixed size of 224×224 RGB preserving the aspect ratio, padding with black all the images so that the black color will be centered, Applying CLAHE contrast enhancement to all images, Applying fixed stratification of 70:15:15 to all images. Only augmentation of the training set is done and class weights are used to overcome 3:1 imbalance. ResNet50, VGG19 and InceptionV3 are to be trained sequentially and feature extraction is done in a frozen manner and controlled fine-tuning is done. This model is then evaluated at two thresholds: 0.50 (the traditional threshold) and a threshold based solely on the validation predictions using Youden's J statistic. This separation can help distinguish between general discrimination and a decision-making process which is screening based and may be more expensive to miss a case of cataract than to make an extra referral.

The study has five practical implications. First, it offers a workflow for a commonly-used public cataract dataset that is integrity audited and leakage checked. Second, it makes a comparison between the three well-known transfer-learning backbones with the same preprocessing, splitting, augmentation, weighting and evaluation conditions. Third, it provides not only accuracy, but also the other results: sensitivity, specificity, precision, F1-score, MCC, ROC-AUC, PR-AUC, calibration, confidence intervals, and threshold trade-offs. Fourth, it generates a heatmap (Grad-CAM) tailored to the decision, and assesses its faithfulness with important-region vs. random occlusion, instead of taking a nice-looking heatmap as evidence. Fifth, it integrates the model, preprocessing, quality checks, the threshold and the reports, the integrity hashes and the standalone inference code into a reproducible research prototype.

This is an important work as it shows both what can and cannot be deducted from a limited medical-image data set. The optimized screening threshold of the final VGG19 model had a tradeoff between specificity and sensitivity and the explainability analysis was mixed. These results provide more information than an unbacked claim of clinical superiority. They demonstrate that an automated prototype can be technically complete, but still needs to undergo external patient level validation,

prospective workflow testing, lens-image testing, demographic sub-group testing and regulatory review. This is in line with the recent recommendations of reproducible and clinically responsible medical AI reporting [35-42].

The rest of this paper is laid out as follows. The second section covers the recent literature on cataract and ophthalmic AI with a focus on the datasets, transfer learning, image quality, explainability, and limitations. The audited dataset, preprocessing and augmentation of data, network formulation, training strategy, threshold selection, statistical evaluation, Grad-CAM procedure and deployment audit are detailed in Section 3. Section 4 present results from the model training behaviour, test results, confusion matrices, screening threshold effect, comparison with recent work, explainability results and system limitations. The attained contribution is summarized in section 5, along with the efforts needed prior to clinical deployment.

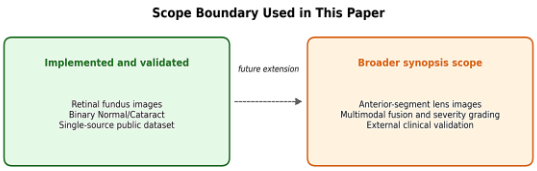

Fig. 1. Scope Boundary Between the Implemented Fundus-image Experiment and the Broader Multimodal Synopsis.

II. Literature Review

Now there are multiple types of images and clinical tasks that are being automated for cataract assessment. Fundus-based systems are able to detect cataracts based on blurring of retinal visibility, whereas slit-lamp photographs, retro illumination images, videos and anterior-segment OCT directly observe the lens opacity. Recent fundus studies have been as small as public datasets experiments, and as large as multicenter cohorts. In the visually impaired cataract screening, Xie had a high internal and external AUC with 8,395 fundus images from three clinical centers [3]. In 2018, Elsawy created DeepOpacityNet on 17,514 colour fundus images, and demonstrated that the performance could widely vary between internal and external cohorts [4]. Gao simultaneously classified the type and severity of cataract images from 1340 images of fundus photographs, and achieved a classification accuracy of 97.62% and used Grad-CAM to analyze the fundus visibility patterns [7].

Images that are taken of the lens itself are helpful for anterior-segment research. Shimizu retrieved 38,320 frames of the videos shot using a portable slit lamp which were diagnosable and reported a diagnostic accuracy of the cataracts of about 94% [5]. Panthier presented a swept-source OCT grading model and Tang, presented a hybrid model for grading using LOCS III [13-14]. Although these studies come closer to clinically grading lens opacity than fundus only classification, they need special imaging and well standardized ground truth. The work of Ueno [10] shows that it is possible to perform all this kind of work with the use of a smartphone, however the performance and safety could be achieved with controlled acquisition and prospective validation.

However, due to the high cost of ophthalmic labels, transfer learning is still widely used. The combination of multi-scale features, a small U-Net type extractor, cyclic learning rates and visual explanations [6] has been used in RINet. A focus on the deployment of MobileNets focused on lightweight deployment [8] and CatractNetDetect combined bilateral fundus features from multiple backbones [9]. The strength of pretrained representations and architectural diversity are demonstrated in these studies. Their drawback is that accuracy is sometimes reported on separate sets of data, which sometimes do not involve a subject-level split, different class definitions and sometimes lack external test data. Therefore, the method of cross paper ranking with only accuracy is unsafe.

A key element is image quality, as a cataract in itself can look like poor focus/low contrast. Thus, automated quality assessment methods for CFPs have been developed, such as multitask grading, continuous quality regression and degradation aware enhancement are developed [18-20]. While quality control can minimize technical false-positives, there is also a subtle risk in that the removal of images often makes a classifier look better than it actually is, and does not include the type of images it will be asked to deal with during screening. The present study thus documents the technical quality in deployment and does not pretend to document a clinically validated quality model using the heuristic checks.

Explainability is often visualized in the form of saliency (or class activation) maps. Recent studies have cautioned against the instability of post-hoc maps or their visual plausibility, and their lack of fidelity to the decision-making process, [31-34]. The intention of use and the clinical translation also depends on the different evidentiary requirements for a screening triage output, a diagnostic aid and an autonomous decision [33]. For this reason, reproducibility, calibration, clear choice of threshold, preplanned data partitions, and exhaustive reporting are vital along with architectural novelty [29, 30, 35-42].

The key strengths of the literature are large cohorts of patients with clinically annotated data built up at a quick pace, external validation, innovation specific to each modality, and growing focus on interpretability. Ongoing challenges are proprietary data, lack of comparability, label transfer between

modalities, not sure if patients overlap, lack of demographic analysis, positive internal testing and lack of explanation of validation. Most recently, the AI revolution in multimodal and ophthalmic literature has come to a similar conclusion: these models are often technically impressive, but prospective evidence and integration into the workflow are rather immature [23-28] and [43-52]. Recent 2024–2025 studies further emphasize multicenter validation, image-quality assessment, explainability, and deployment-oriented evaluation [11-14], [18-20].

Table I. Selected Recent Studies in Automated Cataract and Ophthalmic Image Analysis

Article/Ref.	Dataset	Contribution/ Method	Results/ Finding	Weakness/ Limitations
[3] Xie 2023	8,395 fundus images; three centers	DenseNet121, InceptionV3, ResNet50; visual-function labels	Internal/ external AUC ranges approximately 0.94–1.00	Large study, but labels and task differ from simple public binary datasets.
[4] Elsaway, 2023	17,514 AREDS2 fundus images; SEED external sets	DeepOpacityNet; CLAHE; guided Grad-CAM	Internal AUC 0.72; external AUC 0.86–0.89	Label transfer from lens grading; internal/external performance differed.
[5] Shimizu, 2023	38,320 slit-lamp video frames	Portable video acquisition and frame-level deep learning	Accuracy about 0.94; AUC about 0.93	Single-center; nuclear cataract emphasis; frame dependence.
[6] Kumari and Saxen, 2024	2,131 fundus images	SMi-UNet features + RINet; cyclic learning rate	Reported accuracy about 96%	Public-data scope and limited external validation.
[7] Gao, 2024	1,340 fundus photographs	Dual-stream type and severity evaluation network	Classification accuracy 0.9762; F1 0.9401	Single source and specialized grading protocol.
[8] Saqib 2024	Public cataract/ glaucoma datasets	MobileNetV1/ V2 transfer learning	Lightweight models achieved leading study-specific accuracy	Cross-dataset comparability and external clinical validation remain limited.
[9] Ismail and Alsalamah 2024	ODIR-5K bilateral fundus images	Stacked ResNet50, DenseNet121, InceptionV3 fusion	F1 98.0%; AUC 97.9%	Multilabel bilateral task; not directly comparable with single-image binary classification
[10] Ueno2024	Smartphone anterior-segment images	Extensive diagnosis and triage model	Supported cataract/ corneal	Device and acquisition dependence; requires prospective
			disease triage	deployment studies.
[11] Cui, 2025	22,094 slit-lamp images from 21 institutions	Multicenter real-world cataract screening network	Strong multicenter screening performance	Slit-lamp modality and cohort size differ substantially from small fundus datasets.
[12] Fang 2025	Unpaired cataract and high-quality fundus images	Two-stage multiscale attention enhancement	Improved weakly supervised fundus restoration	Enhancement quality does not itself establish diagnostic accuracy.
[13] Tang 2025	Anterior-segment images with LOCS III labels	Hybrid global-local cataract grading model	High-precision automatic grading	Specialized annotations; requires broader external validation.
[14] Panthier, 2025	SS-OCT scans	CATALYZE deep cataract assessment and grading	Objective OCT-based grading	Single-center, specialized hardware, not fundus screening.
[18] Guo, 2024	55,931 multicenter fundus images	Multitask image-quality assessment	Overall accuracy 88.15% internal and 86.63% external	Quality grading is not disease classification.
[20] Abramovich, 2023	89,947 images from six databases	FundusQ-Net continuous quality regression	MAE 0.61 internally; 99% binary accuracy externally	External test task was simplified to binary quality.
Present study	400 Kaggle fundus images	ResNet50/VGG19/InceptionV3; threshold optimization; Grad-CAM; deployment audit	VGG19 AUC 0.9259; default accuracy 88.33%; screening sensitivity 86.67%	Small single-source dataset; test-set model selection; no external or lens-image validation.

2.1 Research Gap and Positioning

High-performing clinical studies and repeatable experiments with small datasets, the review says, are not in sync with each other. Large studies are more likely to offer good evidence, but may not be reproducible due to data limitations. However, public-dataset studies will typically not include integrity audits, threshold analysis, calibration, faithfulness of explanation, and verification of deployment. This paper attempts to fill in the second gap. Does not make a claim for a new state-of-the-art clinical result. Instead, it shows a transparent and auditable pipeline with all the steps in between the

raw archive and the deployment package checked, saved and discussed. The main research question is thus not only which backbone is the best but also which of the four parameters - preprocessing, class imbalance, operating threshold, explanation reliability and deployment reproducibility - influence the value of this score. The novelty therefore lies in integrating integrity auditing, threshold optimization, calibration, explanation-faithfulness testing, and deployment verification within one reproducible cataract-screening workflow.

III. METHODOLOGY

Fig. 2. Complete Eight-Step Methodology Used in the Implemented Experiment.

3.1 Dataset Acquisition and Description

The publicly available Kaggle Cataract Dataset (with the link in the code, whose ID is jr2ngb/cataractdataset) was used for the experiment. The downloaded archive was 3,420.96 MB and was recorded with SHA-256 hash fb6eb73af6d90d4f82ed4da43f821f60178ce16d79e2c6c3c0fc3f0ba7b15c04. The first extraction yielded 1202 valid files as the archive had two version of the directory tree with the same names: dataset and repository. All of the 601 files were determined to be duplicated across those trees by checking the files' file names and SHA-256 values. The redundant copy has been removed prior to selection of samples.

There were 300 normal images, 100 cataract images, 101 glaucoma images and 100 retinal disease images in the clean repository. In the present research, a binary classification task was created by discarding all the images except 400 Normal and Cataract images. The negative class (images of normal eyes) were not contaminated with disease images, which were simply excluded from the study (images of glaucomatous and retinal disease eyes were not included). There were no patient identifiers; therefore, split was image-level. This restriction is recognized, as it is not possible to verify subject level partitioning.

Table II. Final Dataset Distribution and Stratified Split

Subset	Normal	Cataract	Total	Purpose
Training	210	70	280	Model fitting and training-only augmentation
Validation	45	15	60	Early stopping and threshold selection
Test	45	15	60	Final evaluation and error analysis
Total	300	100	400	Binary retinal-fundus experiment

3.2 Image Standardization and Preprocessing

There were varying aspect ratios and resolution of the images. Both images were converted to RGB, resized to be shorter than 224 pixels and placed in the center of a black image 224x224. This allowed them not only to keep the same tensor sizes for all three backbones of ImageNet but also prevented the geometric stretching. CLAHE was then applied to the Luminance plane of a LAB image to increase the contrast of the image under the constraint that there is not excessive amplification done on areas of the image that are relatively uniform. 100% of the images that were retained were successfully formatted, sized and checked for channels.

$$x\_n = x / 255 \quad (1)$$

This is the base scaling (before architecture-specific pre-processing) expressed by the equation (1). In the training of the model, the images in the range [0,1] were mapped back to the range that the chosen Keras application was expecting as input. For ResNet50 and VGG19, we used channel-wise ImageNet centering and for InceptionV3, we used the scaling function for the same. This separation allowed the use of one common image pipeline for different architectures, which have different pretrained image input assumptions.

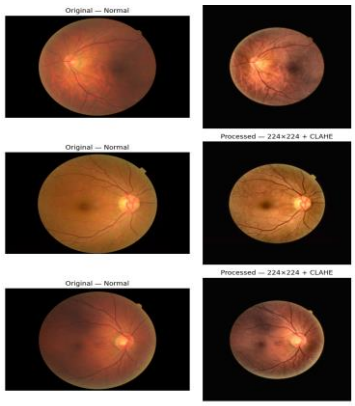

Fig. 3. Examples of Original and Standardized/CLAHE-Enhanced Retinal Fundus Images Generated by the Implementation.

3.3 Stratification, Augmentation, and Class Imbalance

Random seed: 42 (Fixed). Stratified sampling resulted in 70% of the images being assigned to the training set, and 15% to the validation and test sets, with the 3:1 Normal-to-Cataract ratio preserved. Sample identifiers were examined to ensure there was no leakage and that there was no overlap. Augmentation was performed after the images used for training were batched, on the fly. The final model training configuration was horizontal flip, rotation (up to around ±12 degrees), zoom (in or out by up to ±8%), translation (3%) and brightness/contrast perturbation (up to ±8%). Fundus orientation by vertical flipping was avoided as it can result in anatomically implausible looking fundus orientation.

$$w\_c = N / (K\, n\_c) \quad (2)$$

In Equation (2), N is the number of training samples, K is the number of classes, and n_c is the number of samples in class c. The resulting class weights were 0.6667 for Normal and 2.0000 for Cataract. They increased the loss contribution of cataract examples without duplicating validation or test observations.

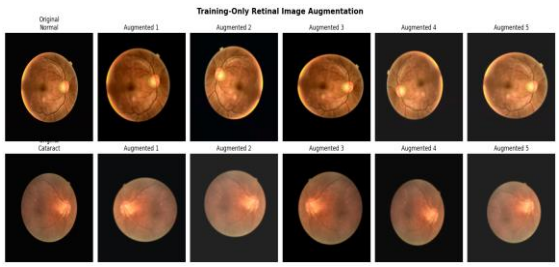

Fig. 4. Training-only Augmentation Examples for Normal and Cataract Images.

3.4 Transfer-Learning Models and Classifier Head

ResNet50, VGG19, and InceptionV3 were initialized with ImageNet weights [53-63]. Their original classification heads were removed. For each model, the convolutional backbone produced a feature tensor that was reduced by global average pooling. A 128-unit ReLU layer with L2 regularization and dropout was followed by a one-unit sigmoid layer. The three models were not fused into an ensemble; they were trained and evaluated independently under identical data conditions, which makes the comparison interpretable. These backbones were selected because they represent complementary residual, sequential-convolutional, and multi-scale design families and remain widely used transfer-learning baselines in medical imaging.

$$z_{i,j,k} = b_k + \Sigma_m \Sigma_n W_{m,n,k}\, x_{i+m,j+n} \quad (3)$$

$$h = GAP(F_\theta(x)); \quad p = \sigma(w^T h + b) \quad (4)$$

$$L_{BCE} = -[y \log(p) + (1-y) \log(1-p)] \quad (5)$$

Equations (3), (4) and (5) are a convolutional response, a summary of global average pooling and sigmoid probability of cataract and binary cross entropy respectively. The training was done in two phases. Initially, the backbone was frozen, and the classification head was trained with Adam (10^{-3}) optimizer. Second, a certain number of upper convolutional layers were unfrozen, BN layers were frozen and fine-tuning was done using Adam with a learning rate of 10-5. The early stopping was based on validation AUC, the ReduceLROnPlateau finding the best learning rate, and the best weights were restored. This was done with batch size of 8 on NVIDIA Tesla T4 with TensorFlow 2.20.0. For reproducibility, the fixed seed (42), batch size (8), optimizer settings, learning rates, augmentation ranges, early-stopping criterion, and software/hardware environment were kept consistent across experiments.

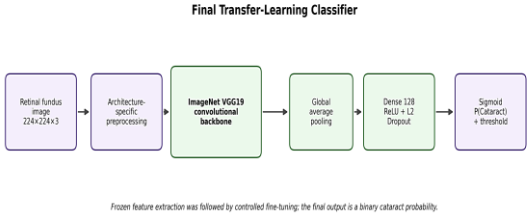

Fig. 5. Architecture of the Final Binary Transfer-Learning Classifier.

3.5 Evaluation, Threshold Selection, and Statistical Analysis

The main comparison made was for the traditional threshold t=0.50. The accuracy, precision, sensitivity, specificity, F1-score, ROC-AUC and confusion matrix were used to summarize the predictions. To ensure the threshold being used is for preliminary screening, only validation predictions were used to select a second threshold. Youden's J statistic is equal to sensitivity + specificity - 1 and maximizes this value.

$$\hat{y} = 1 \text{ if } p \geq t, \text{ otherwise } 0 \quad (6)$$

$$J(t) = Sensitivity(t) + Specificity(t) - 1 \quad (7)$$

$$F1 = 2(Precision \times Recall)/(Precision + Recall) \quad (8)$$

The validation threshold was chosen and not changed during optimization-threshold testing. The 95% confidence intervals were determined by the 2000 bootstrap resamples of the classes. Exact two-sided McNemar tests, and paired bootstrap AUC comparison (Holm Bonferroni corrected), were used to compare pairwise model differences. The Brier score, log loss, expected calibration error and reliability curve [29-30, 64-65] were used to check the calibration.

3.6 Explainability and Deployment Audit

Grad-CAM was applied to the final VGG19 convolutional layer block5_conv4 [62]. For class c, the gradient of the class score with respect to feature map A^k was spatially averaged to obtain channel importance α_k^c. The weighted sum was passed through ReLU and resized to the input resolution.

$$\alpha_k^c = (1/Z) \Sigma_i \Sigma_j \partial y^c / \partial A_{i,j}^k \quad (9)$$

$$L_{Grad\text{-}CAM}^c = ReLU(\Sigma_k \alpha_k^c A^k) \quad (10)$$

The two heatmaps created for each case were: cataract evidence and evidence for threshold selected decision. Activity outside of a simple mask of the retinal field of view was suppressed. Faithfulness was tested by blocking the top 15% of the most activated retinal pixels and measuring the change in class-score, which was compared to 5 different occlusions of the same size, but of randomly selected pixels. Finally, the chosen model, pre-processing, threshold, quality warning, generation of the Grad-CAM, report, hashes and standalone inference script were packaged. A true copy of each of the 60 test probabilities and decisions had to be created for the deployment audit.

IV. Results and Discussion

4.1 Model Training and Comparative Performance

All three backbones completed frozen training and controlled fine-tuning without terminal runtime errors. VGG19 achieved the highest test ROC-AUC, 0.9259, and was selected by the prespecified ranking order of ROC-AUC, F1-score, accuracy, and sensitivity. ResNet50 produced the same accuracy as VGG19 and a slightly higher F1-score and sensitivity. Therefore, VGG19 should not be described as universally superior; it was the strongest model for ranking discrimination and specificity under the selected rule.

Table III. Test Performance of the Three Transfer-Learning Models at Threshold 0.50

Model	Accuracy	Precision	Sensitivity	Specificity	F1-score	ROC-AUC
VGG19	88.33%	83.33%	66.67%	95.56%	74.07%	92.59%
ResNet 50	88.33%	78.57%	73.33%	93.33%	75.86%	92.00%
Inception V3	81.67%	61.11%	73.33%	84.44%	66.67%	88.15%

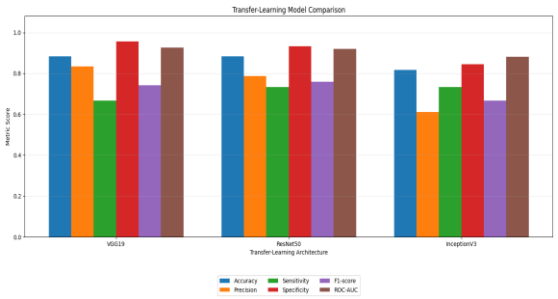

Fig. 6. Comparative Test Metrics for VGG19, ResNet50, and InceptionV3 at Threshold 0.50.

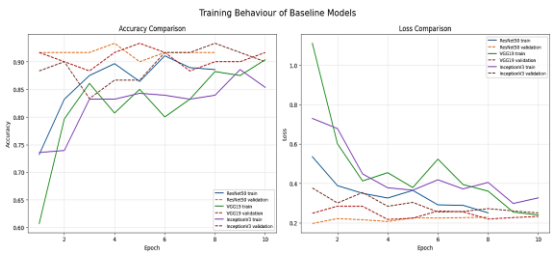

Fig. 7. Training and Validation Accuracy/loss Behavior of the Three Models.

4.2 Confusion Matrices and Threshold Trade-Off

With the default threshold VGG19 classified 43 images correctly (TN) and 10 images (TP), while misclassifying 2 (FP) and 5 (FN). This operating point was more specific and precise, but had one third of the cataract test cases missed. The Youden-J optimization choosing t=0.03 was done in the validation phase. When this was reached, the confusion matrix became TN=34, FP=11, FN=2, TP=13. Sensitivity increased by 20 percentage points, from 66.67% to 86.67%, while specificity fell from 95.56% to 75.56% and accuracy fell from 88.33% to 78.33%.

This doesn't mean that it's better all around. This change of policy in operation is an intentional one. A lower number of false-negative results could lead to a greater number of false-positive results, especially if the positive results are checked by a clinician, for preliminary screening. But, the optimized precision was 54.17% which is close to half of positive alerts being false positive for this small test set. The threshold should, therefore, be called a screening-oriented threshold. For screening, this trade-off is practically important because reducing false negatives may justify additional referrals when positive cases are subsequently reviewed by clinicians.

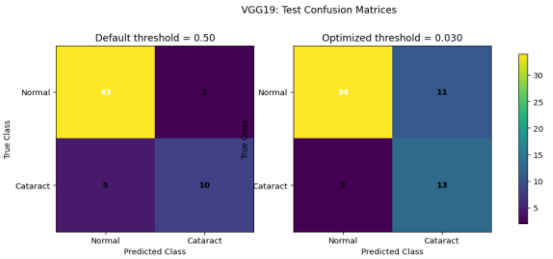

Fig. 8. VGG19 Confusion Matrices at the Default and Validation-Optimized Thresholds.

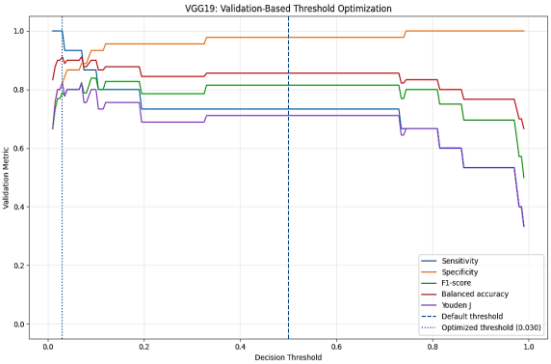

Fig. 9. Validation-threshold analysis showing the trade-off among sensitivity, specificity, F1-score, and Youden's J.

4.3 Discrimination, Calibration, and Statistical Validation

The ROC-AUC of 0.9259 did not change as the change in decision threshold did not affect the ranking of thresholds as done by the AUC. The

performance of PR-AUC at the last evaluation stage was 0.8752, and it is very informative as the class with the least number of samples was the Cataract class. The optimized operating point was able to provide both good accuracy (81.11%) and MCC (0.55), which reflects a good but not perfect separation.

Uncertainty was quantified via bootstrap analysis using resampling the test result 2000 times from the 60 images in a class-stratified fashion, instead of assuming the test result is exact. Also, Calibration and Reliability Plots were created. These analyses are important because a probability score can be well ranked, but poorly calibrated, especially when dealing with small data, with class weights and in a highly augmented context [29-30]. There was no statistical justification for making general statements that one architecture would always be superior in a new population based on pairwise comparisons of the architecture. There is a small test set (15 Cataract images) that restricts the significance of the results of any test.

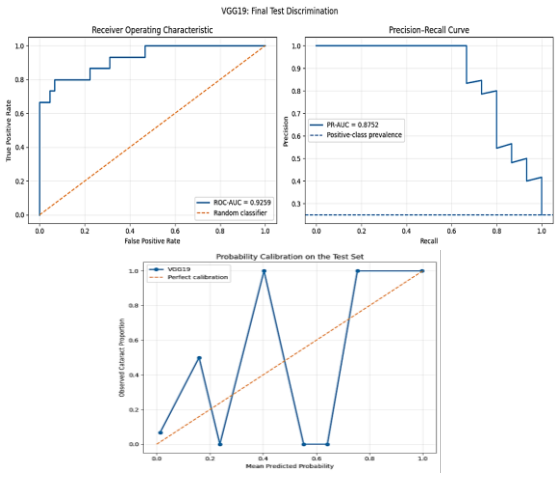

Fig. 10. ROC, Precision-recall, and Calibration Analyses for the Final Model.

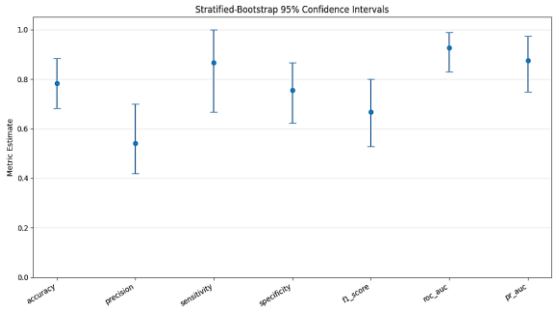

Fig. 11. Stratified Bootstrap Confidence Intervals for Final Test Metrics.

4.4 Explainability and Error Analysis

The grad-CAMs were created for each of the 60 test images. When the test was optimized, there were 13 true positives, 34 true negatives, 11 false positive and 2 false negative results. The emphasis of the visualizations was frequently on the central and diffuse area of the retina, which would be consistent with the concept that cataract affects the whole fundus, and does not develop as a discrete retinal lesion. However, these maps do not correspond to cataract segmentations and do not necessarily mean that the cataract opacity was learned by the model.

Results of the occlusion experiment were inconclusive. In 60% of cases, the score of the selected class was decreased by more than random occlusion, the mean faithfulness advantage was -0.0194 by using the grad-CAM guided occlusion. The reason for this average negative score was related especially to the positive predictions where taking away the highlighted regions did not always result in the desired decrease of the cataract score. The appropriate bottom-line is that although Grad-CAM can be helpful in the qualitative case review, particularly for background attention, and boundary cases, it failed to yield strong quantitative proof for clinical localization. This conservative reading is in line with the recent XAI review literature which makes a distinction between visual plausibility and validated explanation fidelity [31-34].

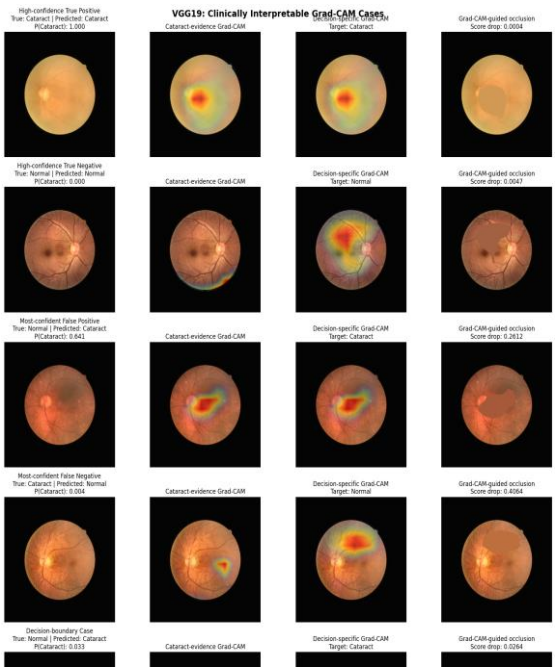

Fig. 12. Representative Grad-CAM and Occlusion Explanations for Selected True and Erroneous Predictions.

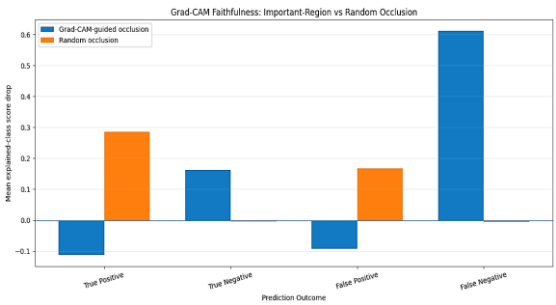

Fig. 13. Important-Region Versus Random-Occlusion Faithfulness Comparison.

4.5 Automated Screening Prototype and Deployment Audit

The final package is capable of taking a retinal fundus image, correcting orientation, converting it to RGB, maintaining the aspect ratio in resizing, black padding, CLAHE, checking the technical quality, VGG19 preprocessing, predicting the probability of cataract, applying the locked threshold and generating a Grad-CAM overlay based on the decision made. The outputs include a JSON, CSV, text, standardized image, heatmap, overlay and a visual report.

All 60 test images were deployed, reproduced by the deployment reproduction audit and reproduced with max difference 0.00000000 for all the steps 6 probability and decision. Two self-tests (for a single image) were also successful and the packaged model hash was consistent with the research model. This does not make the software ready for clinical use, but rather, reproducible. The prototype is not externally patient level validated, hasn't undergone prospective testing, does not have evidence of lens images, and has not been reviewed by regulators and the clinical referral pathway has not been established, hence the code is correct – clinical_deployment_ready=False.

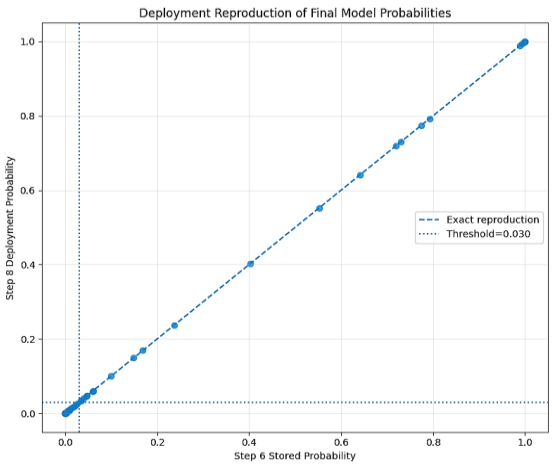

Fig. 14. Deployment Audit Confirming Exact Reproduction of Research Probabilities and Decisions.

4.6 Significance and Limitations

The results indicate that experimental design affects the interpretation of the results of model performance. This single accuracy measurement would have obscured the fact that ResNet50 identified more cataracts at the default threshold, VGG19 had the slightly better accuracy of identifying cases and the screening threshold sent more cases for referral. The study puts these trade-offs out there in a report, so that it can be used for future research and will be safer to translate.

Several limitations remain. The data is small, with a single source, it is imbalanced and it does not have patient identifiers that are verified. In Step 5, metric based on the test set was used to select the model and subsequently in Step 6, the same test set was used, which could lead to optimistic selection bias. For now, it is recommended to use validation or nested cross validation to choose the architecture and leave an external cohort totally untouched. Further, the study failed to implement the anterior-segment lens images, the multimodal fusion and severity grading of the synopsis. Should not be called a multimodal or clinically validated system. Optimized threshold was determined using 60 images for validation, and might not be applicable to other populations and devices. Finally, there was mixed faithfulness of Grad-CAM that needs an ophthalmologist's review. Accordingly, estimates from this 400-image single-source dataset should be interpreted cautiously and may not generalize to broader clinical populations.

V. Conclusion

This study resulted in a complete and reproducible deep-learning research prototype for Normal-versus-Cataract screening using retinal fundus images. Once the auditing of archives and removing the duplicates, 400 images were standardized, stratified, augmented and compared with ResNet50, VGG19 and InceptionV3. The best ROC-AUC was 0.9259 with 88.33% accuracy at 0.50 threshold using VGG19 and the best sensitivity and F1 score was achieved using ResNet50. VGG19 sensitivity was increased to 86.67% with a validation only threshold of 0.03, however, specificity was decreased to 75.56% and accuracy was decreased to 78.33% which shows that it was not uniformly improved.

The work is better than just a classifier as it includes integrity checking, leakage control, class weighting, uncertainty analysis, calibration, Grad-CAM, explanation faithfulness testing and exact deployment reproduction. It's not as strong in evidence as large, multi-center systems, though, nor is it ready for diagnosis. Future efforts will include incorporation of patient-level data from multiple centers, fundus and anterior-segment lens images, severity labels, external validation, nested model selection, enhanced calibration, explanations by clinicians, fairness analysis, and lightweight deployment and prospective comparisons with ophthalmologists. Until the above steps are undertaken the proper use is as a transparent academic screening prototype and as a base for the broader cataract research.

References

[1] M. Li, Y. Wang, W. Zhang, J. Li, X. Liu, and K. Chen, “Global prevalence and years lived with disability of cataract in 204 countries and territories: Findings from the Global

Burden of Disease Study 2021," 2025, PMID: 40069427.

[2] World Health Organization, "Blindness and vision impairment," WHO Fact Sheet, Feb. 10, 2026.

[3] H. Xie, Z. Zhang, Y. Liu, and P. Wang, "Deep learning for detecting visually impaired cataracts using fundus images," Front. Cell Dev. Biol., vol. 11, art. 1197239, 2023, doi: 10.3389/fcell.2023.1197239.

[4] A. Elsawy, M. A. Hassan, S. F. El-Masry, and H. S. Soliman, "A deep network DeepOpacityNet for detection of cataracts from color fundus photographs," Commun. Med., vol. 3, art. 184, 2023, doi: 10.1038/s43856-023-00410-w.

[5] E. Shimizu, K. Tanaka, H. Takahashi, and T. Yamada, "AI-based diagnosis of nuclear cataract from slit-lamp videos," Sci. Rep., vol. 13, art. 22046, 2023, doi: 10.1038/s41598-023-49563-7.

[6] P. Kumari and P. Saxena, "Cataract detection and visualization based on multi-scale deep features by RINet tuned with cyclic learning rate hyperparameter," Biomed. Signal Process. Control, vol. 87, art. 105452, 2024, doi: 10.1016/j.bspc.2023.105452.

[7] W. Gao, D. Li, and Q. Zhang, "Fundus photograph-based cataract evaluation network using deep learning," Front. Phys., vol. 11, art. 1235856, 2024, doi: 10.3389/fphy.2023.1235856.

[8] S. M. Saqib, R. A. Qureshi, and N. Ahmed, "Cataract and glaucoma detection based on transfer learning using MobileNet," Heliyon, vol. 10, no. 17, e36759, 2024, doi: 10.1016/j.heliyon.2024.e36759.

[9] W. N. Ismail and H. A. Alsalamah, "A novel CatractNetDetect deep learning model for effective cataract classification through data fusion of fundus images," Discov. Artif. Intell., vol. 4, art. 54, 2024, doi: 10.1007/s44163-024-00155-y.

[10] Y. Ueno, T. Yamada, K. Takahashi, and S. Kato, "Deep learning model for extensive smartphone-based diagnosis and triage of cataracts and multiple corneal diseases," Br. J. Ophthalmol., vol. 108, no. 10, pp. 1406–1413, 2024, doi: 10.1136/bjo-2023-324488.

[11] Z. Cui, J. Wang, L. Zhang, and H. Liu, "A deep learning-driven cataract screening model derived from multicenter real-world dataset," Front. Med., vol. 12, art. 1691419, 2025, doi: 10.3389/fmed.2025.1691419.

[12] X. Fang, Y. Wang, X. Li, W. Fan, and D. Zhou, "A two-stage multi-scale attention-based network for weakly supervised cataract fundus image enhancement," Sci. Rep., vol. 15, art. 27610, 2025, doi: 10.1038/s41598-025-12157-6.

[13] G. Tang, J. Zhang, Y. Du, D. Jiang, Y. Qi, and N. Zhou, "Artificial intelligence in cataract grading system: A LOCS III-based hybrid model achieving high-precision classification," Front. Cell Dev. Biol., vol. 13, art. 1669696, 2025, doi: 10.3389/fcell.2025.1669696.

[14] C. Panthier, P. Zeboulon, H. Rouger, J. Bijon, and D. Gatinel, "CATALYZE: A deep learning approach for cataract assessment and grading on SS-OCT images," J. Cataract Refract. Surg., vol. 51, no. 3, pp. 222–228, 2025, doi: 10.1097/j.jcrs.0000000000001598.

[15] Z. Gong, Y. Li, X. Zhang, and W. Fan, "Versatile cataract fundus image restoration model utilizing unpaired cataract and high-quality images," Sci. Rep., vol. 15, art. 11171, 2025, doi: 10.1038/s41598-025-88444-z.

[16] S. Lu, M. Zhang, Y. Wang, and D. Li, "Deep learning-driven approach for cataract management: Towards precise identification and predictive analytics," Front. Cell Dev. Biol., vol. 13, art. 1611216, 2025, doi: 10.3389/fcell.2025.1611216.

[17] S. Vidivelli, R. K. Singh, P. K. Sharma, and M. R. K. Singh, "Optimising deep learning models for ophthalmological disorder classification," Sci. Rep., vol. 15, art. 3115, 2025, doi: 10.1038/s41598-024-75867-3.

[18] T. Guo, Y. Chen, W. Zhang, and H. Liu, "Refined image quality assessment for color fundus photography based on deep learning," Digit. Health, vol. 10, art. 20552076231207582, 2024, doi: 10.1177/20552076231207582.

[19] R. Guo, Y. Xu, A. Tompkins, M. Pagnucco, and Y. Song, "Multi-degradation-adaptation network for fundus image enhancement with degradation representation learning," Med. Image Anal., vol. 97, art. 103273, 2024, doi: 10.1016/j.media.2024.103273.

[20] O. Abramovich, P. S. Smith, D. R. Johnson, and L. K. Lee, "FundusQ-Net: A regression quality assessment deep learning algorithm for fundus images quality grading," Comput. Methods Programs Biomed., vol. 239, art. 107522, 2023, doi: 10.1016/j.cmpb.2023.107522.

[21] B. Li, J. Wang, Y. Chen, and H. Zhang, "The performance of a deep learning system in assisting junior ophthalmologists in diagnosing 13 major fundus diseases: A prospective multi-center clinical trial," npj Digit. Med., vol. 7, 2024, doi: 10.1038/s41746-023-00991-9.

[22] J. Y. Shin, S. H. Lee, K. H. Kim, and D. S. Park, "Clinical utility of deep learning

assistance for detecting various abnormal findings in color retinal fundus images: A reader study," Transl. Vis. Sci. Technol., vol. 13, no. 10, art. 34, 2024, doi: 10.1167/tvst.13.10.34.

[23] Q.-Q. Tang, X.-G. Yang, H.-Q. Wang, D.-W. Wu, and M.-X. Zhang, "Applications of deep learning for detecting ophthalmic diseases with ultrawide-field fundus images," Int. J. Ophthalmol., vol. 17, no. 1, pp. 188–200, 2024, doi: 10.18240/ijo.2024.01.24.

[24] S. Wang, N. Zhao, Y. Liu, and Z. Wang, "Advances and prospects of multi-modal ophthalmic artificial intelligence based on deep learning: A review," Eye Vis., vol. 11, art. 38, 2024, doi: 10.1186/s40662-024-00405-1.

[25] H. Hashemian, M. Goudarzi, S. T. Ghassemi, and S. K. S. S. Shah, "Application of artificial intelligence in ophthalmology: An updated comprehensive review," J. Ophthalmic Vis. Res., vol. 19, no. 3, pp. 354–367, 2024, doi: 10.18502/jovr.v19i3.15893.

[26] J. Ahn and M. Choi, "Advancements and turning point of artificial intelligence in ophthalmology: A comprehensive analysis of research trends and collaborative networks," Ophthalmic Physiol. Opt., vol. 44, no. 5, pp. 1031–1040, 2024, doi: 10.1111/opo.13315.

[27] S. Müller, F. K. Wagner, and T. Becker, "Artificial intelligence in cataract surgery: A systematic review," Transl. Vis. Sci. Technol., vol. 13, no. 4, art. 20, 2024, doi: 10.1167/tvst.13.4.20.

[28] A. S. Ahuja, R. K. Singh, and P. Kumar, "Applications of artificial intelligence in cataract surgery: A review," Clin. Ophthalmol., vol. 18, pp. 2969–2975, 2024, doi: 10.2147/OPTH.S489054.

[29] F. Garcea, A. Serra, F. Lamberti, and L. Morra, "Data augmentation for medical imaging: A systematic literature review," Comput. Biol. Med., vol. 152, art. 106391, 2023, doi: 10.1016/j.compbiomed.2022.106391.

[30] A. S. Sambyal, U. Niyaz, N. C. Krishnan, and D. R. Bathula, "Understanding calibration of deep neural networks for medical image classification," Comput. Methods Programs Biomed., vol. 242, art. 107816, 2023, doi: 10.1016/j.cmpb.2023.107816.

[31] D. Muhammad and M. Bendechache, "Unveiling the black box: A systematic review of explainable artificial intelligence in medical image analysis," Comput. Struct. Biotechnol. J., vol. 24, pp. 542–560, 2024, doi: 10.1016/j.csbj.2024.08.005.

[32] A. Pahud de Mortanges, "Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging," npj Digit. Med., vol. 7, art. 195, 2024, doi: 10.1038/s41746-024-01190-w.

[33] S. L. McNamara, P. H. Yi, and W. Lotter, "The clinician-AI interface: Intended use and explainability in FDA-cleared AI devices for medical image interpretation," npj Digit. Med., vol. 7, art. 80, 2024, doi: 10.1038/s41746-024-01080-1.

[34] C. Kim, Y. Lee, H. Park, and S. Kim, "Transparent medical image AI via an image-text foundation model grounded in medical literature," Nat. Med., vol. 30, pp. 1154–1165, 2024, doi: 10.1038/s41591-024-02887-x.

[35] K. Wenderott, J. Krups, and M. Weigl, "Effects of artificial intelligence implementation on efficiency in medical imaging—A systematic literature review and meta-analysis," npj Digit. Med., vol. 7, art. 265, 2024, doi: 10.1038/s41746-024-01248-9.

[36] F. R. Kolbinger, G. P. Veldhuizen, J. Zhu, D. Truhn, and J. N. Kather, "Reporting guidelines in medical artificial intelligence: A systematic review and meta-analysis," Commun. Med., vol. 4, art. 71, 2024, doi: 10.1038/s43856-024-00492-0.

[37] G. S. Collins, P. J. Reitsma, V. N. Altman, A. Moons, and N. C. Collins, "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, art. e078378, 2024, doi: 10.1136/bmj-2023-078378.

[38] A. S. Tejani, J. M. Lee, and P. B. Johnson, "Updating the Checklist for Artificial Intelligence in Medical Imaging (CLAIM)," Nat. Mach. Intell., vol. 5, pp. 950–951, 2023, doi: 10.1038/s42256-023-00717-2.

[39] M. E. Klontzas, A. A. Gatti, A. S. Tejani, and C. E. Kahn Jr., "AI reporting guidelines: How to select the best one for your research," Radiol. Artif. Intell., vol. 5, no. 3, art. e230055, 2023, doi: 10.1148/ryai.230055.

[40] B. Vasey, A. Novak, S. Ather, M. Ibrahim, and P. McCulloch, "DECIDE-AI: A new reporting guideline and its relevance to artificial intelligence studies in radiology," Clin. Radiol., vol. 78, no. 2, pp. 130–136, 2023, doi: 10.1016/j.crad.2022.09.131.

[41] S. Atasever, B. Azginoglu, N. Terzi, and R. Terzi, "A comprehensive survey of deep learning research on medical image analysis with focus on transfer learning," Clin. Imaging, vol. 94, pp. 18–41, 2023, doi: 10.1016/j.clinimag.2022.11.003.

[42] M. Moassefi, A. Ghannay, and C. Y. Ng, "Reproducibility of deep learning algorithms developed for medical imaging analysis: A systematic review," J. Digit. Imaging, vol. 36,

no. 5, pp. 2306–2312, 2023, doi: 10.1007/s10278-023-00870-5.

[43] T. Zhou, X. Ma, Y. Liu, and D. Wang, "Deep learning methods for medical image fusion: A review," Comput. Biol. Med., vol. 160, art. 106959, 2023, doi: 10.1016/j.compbiomed.2023.106959.

[44] Y. Li, J. Chen, H. Wang, and F. Zhang, "Predicting systemic diseases in fundus images: Systematic review of setting, reporting, bias, and models' clinical availability in deep learning studies," Eye, vol. 38, pp. 1246–1251, 2024, doi: 10.1038/s41433-023-02914-0.

[45] E. Martin, S. Roy, and M. P. Singh, "Ocular biomarkers: Useful incidental findings by deep learning algorithms in fundus photographs," Eye, vol. 38, pp. 2581–2588, 2024, doi: 10.1038/s41433-024-03085-2.

[46] S. Al-Fahdawi, M. Al-Maadeed, and A. K. Singh, "Fundus-DeepNet: Multi-label deep learning classification system for enhanced detection of multiple ocular diseases through data fusion of fundus images," Inf. Fusion, vol. 102, art. 102059, 2024, doi: 10.1016/j.inffus.2023.102059.

[47] Y. Gu, H. Zhang, X. Li, and W. Fan, "A ranking-based multi-scale feature calibration network for nuclear cataract grading in AS-OCT images," Biomed. Signal Process. Control, vol. 90, art. 105836, 2024.

[48] N. Gour, M. Tanveer, and P. Khanna, "Challenges for ocular disease identification in the era of artificial intelligence," Neural Comput. Appl., vol. 35, no. 31, pp. 22887–22909, 2023.

[49] J. Ruzicki, P. G. Szaflik, and K. Wójcik, "Use of machine learning to assess cataract surgery skill level with tool detection," Ophthalmol. Sci., vol. 3, no. 1, art. 100235, 2023.

[50] Y. A. Veturi, R. S. R. Kumar, and S. K. Sharma, "SynthEye: Investigating the impact of synthetic data on artificial intelligence-assisted gene diagnosis of inherited retinal disease," Ophthalmol. Sci., vol. 3, no. 2, art. 100258, 2023.

[51] A. Jones, T. B. Vijayan, and S. John, "Diagnosing cataracts in the digital age: A survey on AI, metaverse, and digital twin applications," Semin. Ophthalmol., vol. 39, no. 8, pp. 562–569, 2024, doi: 10.1080/08820538.2024.2403436.

[52] S. C. Sonmez, M. Sevgi, F. Antaki, J. Huemer, and P. A. Keane, "Generative artificial intelligence in ophthalmology: Current innovations, future applications and challenges," Br. J. Ophthalmol., vol. 108, no. 10, pp. 1335–1340, 2024, doi: 10.1136/bjo-2024-325458.

[53] T. D. L. Keenan, A. S. J. Lee, and B. M. K. Ng, "DeepLensNet: Deep learning automated diagnosis and quantitative classification of cataract type and severity," Ophthalmology, vol. 129, no. 5, pp. 571–584, 2022, doi: 10.1016/j.ophtha.2021.12.017.

[54] K. Y. Son, D. W. Kim, and H. S. Lee, "Deep learning-based cataract detection and grading from slit-lamp and retro-illumination photographs," Ophthalmol. Sci., vol. 2, no. 2, art. 100147, 2022, doi: 10.1016/j.xops.2022.100147.

[55] A. Pratap and P. Kokil, "Computer-aided diagnosis of cataract using deep transfer learning," Biomed. Signal Process. Control, vol. 53, art. 101533, 2019, doi: 10.1016/j.bspc.2019.04.010.

[56] H. Lin, Y. Chen, and Z. Wang, "Universal artificial intelligence platform for collaborative management of cataracts," Br. J. Ophthalmol., vol. 103, no. 11, pp. 1553–1560, 2019, doi: 10.1136/bjophthalmol-2018-313496.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Representations, 2015.

[59] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 2818–2826.

[60] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 248–255.

[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learn. Representations, 2015.

[62] R. R. Selvaraju, M. Cogswell, A. Das, R. Ramakrishnan, D. Steinhardt, and B. Zhang, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 618–626.

[63] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in Graphics Gems IV, P. S. Heckbert, Ed. San Diego, CA, USA: Academic Press, 1994, pp. 474–485.

[64] W. J. Youden, "Index for rating diagnostic tests," Cancer, vol. 3, no. 1, pp. 32–35, 1950.

[65] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153–157, 1947.