medRxiv preprint doi: https://doi.org/10.64898/2026.03.02.26347127; this version posted March 3, 2026. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

CausalFund: Causality-Inspired Domain Generalization in Retinal Fundus Imaging for Low-Resource Screening

Min Shi, PhD[1], Hao Zheng, PhD[1], Raju Gottumukkala, PhD[2], Nussdorf Jonathan, MD[3], Grayson W. Armstong, MD[4], Lucy Q. Shen, MD[4], Mengyu Wang, PhD[5]

[1] School of Computing and Informatics, University of Louisiana at Lafayette, Lafayette, LA, USA
[2] Informatics Research Institute, University of Louisiana at Lafayette, Lafayette, LA, USA
[3] Ochsner Health in Louisiana, New Orleans, LA, USA
[4] Massachusetts Eye and Ear, Harvard Medical School, Boston, MA, USA
[5] Harvard Ophthalmology AI Lab, Schepens Eye Research Institute of Massachusetts Eye and Ear, Harvard Medical School, Boston, MA, USA

Abstract

Early screening for glaucoma and diabetic retinopathy (DR) is critical to prevent irreversible vision loss, yet remains inaccessible to many underserved populations. However, AI models trained on hospital-grade fundus images often generalize poorly to low-cost images acquired with portable devices such as smartphones. We proposed CausalFund, a causality-inspired learning framework for training AI models that enable reliable low-resource screening from easily acquired non-clinical images. CausalFund disentangles disease-relevant retinal features from spurious image factors to achieve domain-generalizable screening across clinical and non-clinical settings. We integrated CausalFund with seven deep learning backbones for glaucoma and DR screening from portable-device fundus images, including lightweight architectures suitable for on-device deployment. Across diverse experimental settings and image quality conditions, CausalFund consistently improved AUC and achieved a more favorable sensitivity-specificity trade-off than conventional deep learning baselines. As a model-agnostic framework, CausalFund could be extended to other diseases and low-resourced scenarios characterized by degraded or non-standard imaging.

Min Shi (min.shi@louisiana.edu)

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.64898/2026.03.02.26347127; this version posted March 3, 2026. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Introduction

Glaucoma and diabetic retinopathy (DR) are leading causes of preventable vision impairment worldwide,[1,2] and both diseases benefit substantially from early detection and timely referral or treatment.[3,4] However, comprehensive eye exams and high-quality retinal imaging infrastructure are often concentrated in well-resourced clinical centers, leaving financially disadvantaged families and individuals in rural areas without routine examinations.[5,6] This access gap motivates reliable and low-cost solutions that can function outside traditional ophthalmology clinics.

In recent years, artificial intelligence (AI)-based approaches are widely utilized to automatically detect glaucoma and DR using retinal imaging such as optical coherence tomography (OCT) and fundus images (**Figure 1a**), which hold significant promise to alleviate clinical resource burdens and enable affordable screening.[7,8] However, existing AI models depend heavily on high-quality retinal images captured from costly imaging devices,[9] such as those from Topcon and Zeiss, typically available in hospital and ophthalmology clinic-based settings. Such domain-dependent models trained on curated clinical datasets may fail to transfer to new acquisition conditions, devices, and populations.[10,11] In fundus imaging, the shift from hospital-grade cameras to portable devices (e.g., smartphones) may introduce changes in field of view, illumination, sharpness, color response, and imaging artifacts,[12,13] which can substantially degrade model reliability at deployment for cost-effective screening. As a result, current AI screening models may not be easily deployed in low-resource environments where hospital-grade fundus cameras are unavailable. This significantly limits the reach of AI screening solutions to broader populations. An effective way to reduce this gap is to collect labeled target-domain data and retrain or fine-tune models, but this is often impractical in low-resource settings where labeled data and consistent acquisition protocols are difficult to obtain. Moreover, conventional training can encourage reliance on spurious factors, such as patient demographics, device signatures, compression artifacts, or quality-related cues, that correlate with labels in the training environment but are unstable across domains.[14–17] This is especially problematic for non-clinical fundus screening, where capture conditions are uncontrolled and image quality can vary substantially.

This work aims to develop robust fundus screening models that generalize from hospital-grade clinical images to non-clinical, low-cost images acquired with smartphones or portable imaging devices (**Figure 1b**). We argue that a key ingredient is to explicitly discourage models from encoding domain-specific nuisance variation, and instead promote representations aligned with causal, disease-relevant retinal structure, such as the cup-to-disc ratio and neuroretinal rim

medRxiv preprint doi: https://doi.org/10.64898/2026.03.02.26347127; this version posted March 3, 2026. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

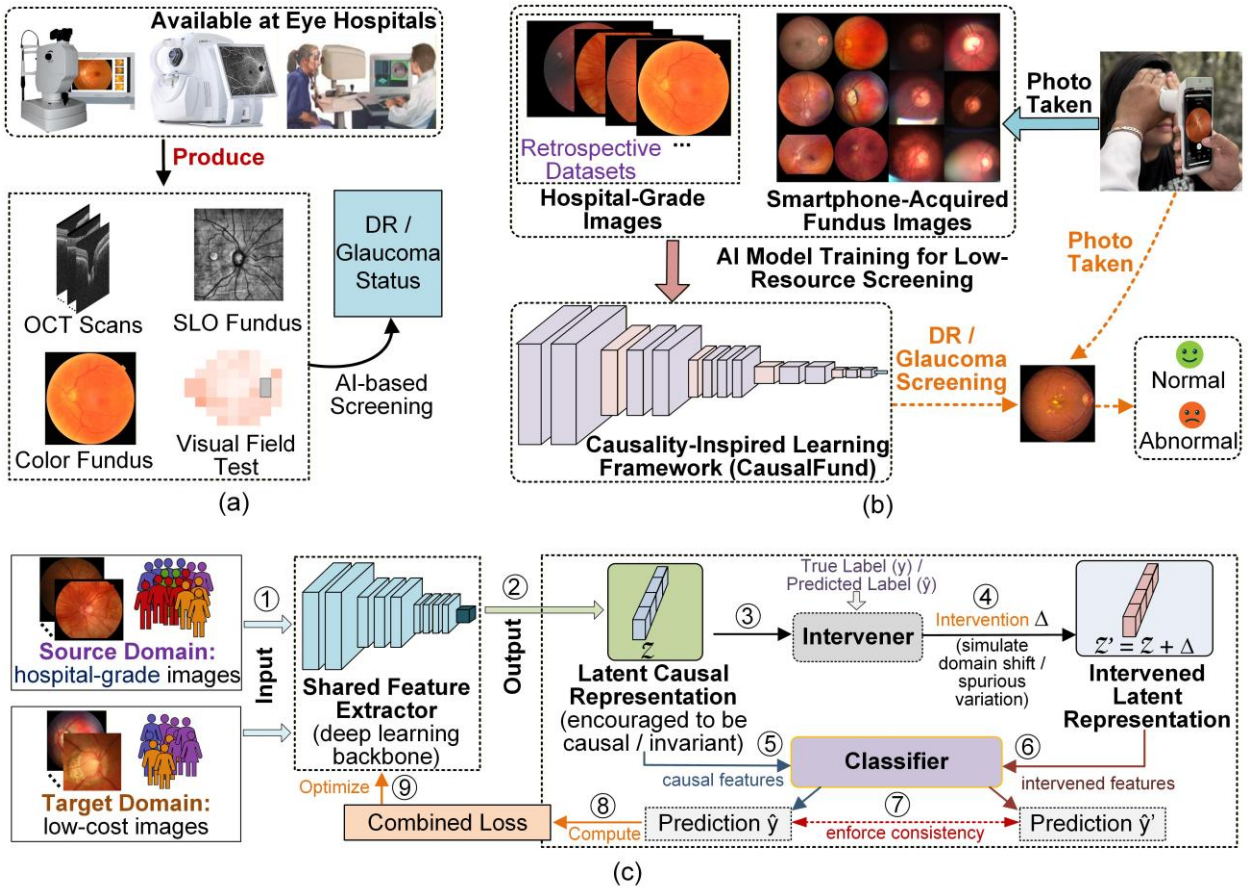


Figure 1: Low-resource screening and the proposed CausalFund framework. (a) Existing AI-based screening using hospital-based imaging devices. (b) Cost-effective AI for low-resource screening. (c) Illustration of the proposed CausalFund framework.

thinning. Our perspective is inspired by broader insights from causal machine learning,[18,19] which suggest that relationships between causal features and labels remain stable across environments, and that learning invariant predictors can substantially improve robustness when spurious correlations shift across domains. Accordingly, we proposed a causality-inspired framework called CausalFund to train cost-effective AI models for low-resource glaucoma and DR screening. CausalFund learns paired causal and spurious representations (**Figure 1c**), in which predictions are explicitly encouraged to rely on disease-causal features while suppressing or disentangling spurious, domain-dependent factors, thereby improving generalization from hospital-grade retinal fundus images to portable-device images. We integrated CausalFund with a range of deep learning backbones, including lightweight architectures suitable for mobile deployment, and evaluated it for both glaucoma and DR screening under multiple experimental settings and image-quality conditions. Across different experimental settings, CausalFund improved screening

medRxiv preprint doi: https://doi.org/10.64898/2026.03.02.26347127; this version posted March 3, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

performance and yielded a more favorable sensitivity-specificity trade-off than conventional AI models, with particularly notable gains in scenarios with severe image-quality degradations.

Results

Dataset Collection

We curated two domain-paired datasets for evaluating domain generalization in retinal fundus screening, including one for glaucoma and another for DR screening. For each task, we paired a hospital-based fundus dataset (e.g., higher-quality camera acquisitions) with a portable or smartphone dataset (e.g., lower-cost acquisitions with greater variability in illumination, focus, and field of view). All datasets were obtained from publicly available sources and used in accordance with their respective licenses and data-use policies. For the glaucoma dataset in hospital domain, we used the Multichannel Glaucoma Benchmark Dataset from Kaggle, which provides fundus images and dataset annotations for glaucoma classification. For the smartphone domain, we used the Brazil Glaucoma (BrG) dataset from Kaggle, which contains retinal images acquired with portable or smartphone-based devices and corresponding glaucoma labels. After preprocessing and patient-level splitting, the hospital glaucoma domain contained 12,316 images and the smartphone glaucoma domain contained 2,000 images (**Table 1**). We treated both datasets as a binary classification task (healthy vs glaucoma) using the dataset-provided annotations.

For the DR dataset in the hospital domain, we used a combined fundus dataset aggregating multiple widely used hospital-grade DR sources (APTOS, DDR, IDRiD, EyePACS, and Messidor) provided on Kaggle. For the smartphone domain, we used the Mobile Brazilian Retinal Dataset (mBRSET) hosted on PhysioNet, which comprises retinal fundus photographs acquired using a handheld portable camera integrated with a smartphone platform and labeled for DR severity using standard DR grading. DR grades follow the conventional five-level scale (0-4: no DR, mild NPDR, moderate NPDR, severe NPDR, proliferative DR). After preprocessing and splitting, the hospital DR domain contained 45,327 images, and the smartphone DR domain contained 4,884 images (**Table 1**).

Glaucoma Screening Results

We compared CausalFund with conventional empirical risk minimization (ERM) based on seven backbone architectures, including five standard models ResNet, DenseNet, EfficientNet, VGG, ViT, and two lightweight models MobileNet and SqueezeNet. CausalFund achieved comparable

medRxiv preprint doi: https://doi.org/10.64898/2026.03.02.26347127; this version posted March 3, 2026. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

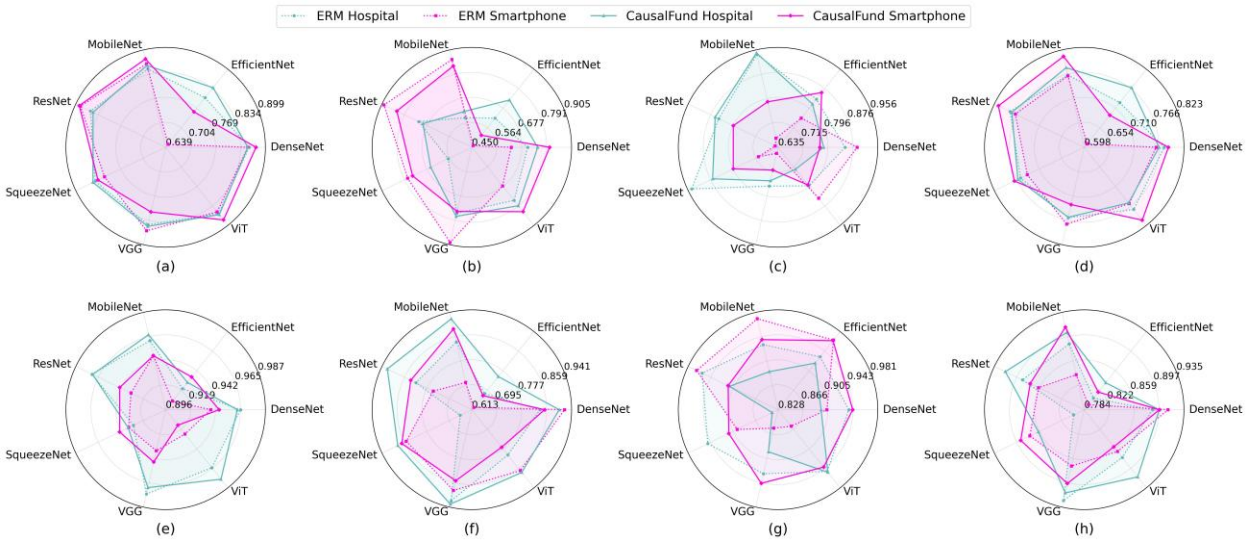


Figure 2: Glaucoma and DR screening results of ERM and CausalFund combined with various backbones, validated on hospital and smartphone-based fundus images. (a) Glaucoma screening AUC. (b) Glaucoma screening Sensitivity. (c) Glaucoma screening Specificity. (d) Sensitivity/specificity averaged results of glaucoma screening. (e) DR screening AUC. (f) DR screening Sensitivity. (g) DR screening Specificity. (h) Sensitivity/specificity averaged results of DR screening.

or higher AUC than ERM in both hospital and smartphone domains (**Figure 2a-d**). In the hospital domain, ERM AUC ranged from 0.804 to 0.860, whereas CausalFund ranged from 0.837 to 0.863. For example, EfficientNet improved from 0.804 to 0.837 ($p < 0.001$) and MobileNet from 0.848 to 0.857 ($p < 0.05$). In the smartphone domain, ERM AUC ranged from 0.649 to 0.885, while CausalFund ranged from 0.757 to 0.889, including significant improvements ($p < 0.001$) of EfficientNet from 0.649 to 0.757, SqueezeNet from 0.816 to 0.835, and ViT from 0.854 to 0.881.

For sensitivity, hospital-domain ERM values ranged from 0.588 to 0.759, compared with 0.617 to 0.791 for CausalFund (**Figure 2b**). Smartphone-domain sensitivity under ERM ranged from 0.460 to 0.895, while CausalFund ranged from 0.520 to 0.830. For specificity, hospital-domain ERM ranged from 0.763 to 0.947 and CausalFund ranged from 0.725 to 0.946 (**Figure 2c**). Smartphone-domain specificity ranged from 0.645 to 0.890 for ERM and from 0.710 to 0.860 for CausalFund. When averaging sensitivity and specificity (**Figure 2d**), for hospital-domain standard models, ERM averages ranged from 0.726 to 0.783 (e.g., 0.783 for ResNet, 0.777 for ViT), compared with 0.758 to 0.782 for CausalFund (e.g., 0.779 for ResNet, 0.758 for ViT). For hospital-domain lightweight models, ERM achieved 0.765 for MobileNet and 0.757 for

medRxiv preprint doi: https://doi.org/10.64898/2026.03.02.26347127; this version posted March 3, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

SqueezeNet, compared with 0.782 and 0.764 for CausalFund. In the smartphone domain, ERM averages for standard backbones ranged from 0.608 to 0.775 (e.g., 0.608 for EfficientNet, 0.770 for ResNet), while CausalFund ranged from 0.690 to 0.813 (e.g., 0.813 for ResNet, 0.808 for ViT). For lightweight models, ERM averaged 0.763 for MobileNet and 0.740 for SqueezeNet, compared with 0.808 and 0.773 for CausalFund.

Diabetic Retinopathy Screening Results

In the hospital domain, ERM AUC ranged from 0.921 to 0.975, whereas CausalFund ranged from 0.928 to 0.977 (**Figure 2e**). For example, EfficientNet improved from 0.921 to 0.928 ($p < 0.05$), DenseNet from 0.965 to 0.962 (comparable), and ViT from 0.964 to 0.977 ($p < 0.05$). In the smartphone domain, ERM AUC ranged from 0.907 to 0.947, while CausalFund ranged from 0.914 to 0.947 (**Figure 2e**), including improvements for DenseNet from 0.938 to 0.945 ($p < 0.05$) and EfficientNet from 0.907 to 0.935 ($p < 0.001$).

For sensitivity (**Figure 2f**), hospital-domain ERM values ranged from 0.655 to 0.921, compared with 0.753 to 0.931 for CausalFund. Smartphone-domain sensitivity ranged from 0.623 to 0.918 under ERM and from 0.672 to 0.885 under CausalFund. For specificity (**Figure 2g**), hospital-domain ERM ranged from 0.929 to 0.958, while CausalFund ranged from 0.888 to 0.951. Smartphone-domain specificity ranged from 0.857 to 0.971 for ERM and from 0.912 to 0.964 for CausalFund. When averaging sensitivity and specificity (**Figure 2h**), hospital-domain ERM values ranged from 0.802 to 0.925, while CausalFund ranged from 0.836 to 0.914. In the smartphone domain, ERM averages ranged from 0.794 to 0.911, compared with 0.818 to 0.912 for CausalFund. For lightweight models specifically, hospital-domain ERM achieved 0.886 for MobileNet and 0.802 for SqueezeNet, compared with 0.904 and 0.861 for CausalFund. In the smartphone domain, ERM averaged 0.838 for MobileNet and 0.875 for SqueezeNet, compared with 0.912 and 0.891 for CausalFund.

Impact of Smartphone Image Quality Degradation on Screening Results

Glaucoma screening results showed consistent gains from CausalFund over ERM across image-quality degradation levels (**Figure 3**), with the strongest impacts to smartphone-domain screening. Mean improvements in AUC across all seven backbones were positive in all six settings (from 0.009 to 0.059, $p < 0.05$), peaking under severe degradation level. Improvements of averaged sensitivity-specificity Avg(Sens, Spec) were also positive in all six settings (0.008 to 0.054, $p < 0.05$), again highest for severe degradation level. Sensitivity gains were generally favorable on smartphone screening (mild 0.062, moderate 0.058, severe 0.045, $p < 0.001$), while specificity

medRxiv preprint doi: https://doi.org/10.64898/2026.03.02.26347127; this version posted March 3, 2026. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

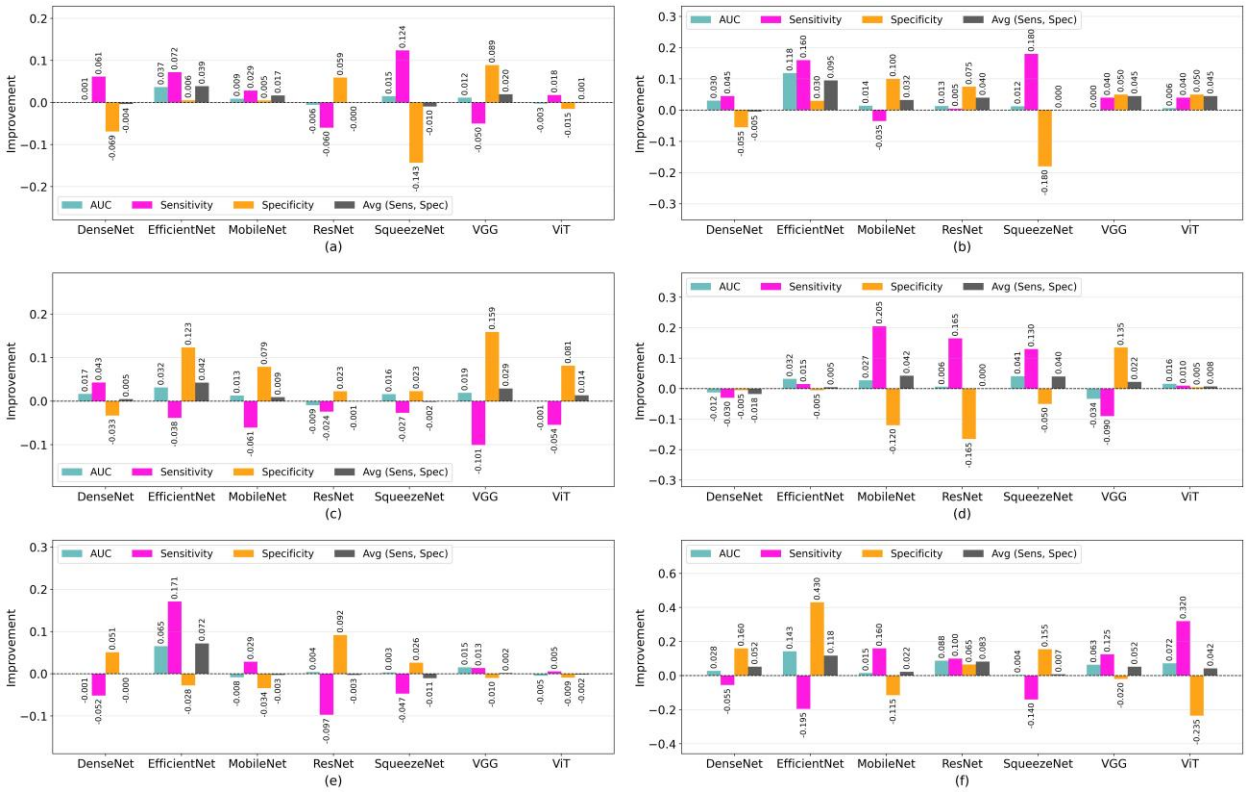


Figure 3: Glaucoma screening improvements of CausalFund over ERM after applying different levels (mild, moderate and severe) of degradation on smartphone images. (a) Hospital screening improvements after mild degradation. (b) Smartphone screening improvements after mild degradation. (c) Hospital screening improvements after moderate degradation. (d) Smartphone screening improvements after moderate degradation. (e) Hospital screening improvements after severe degradation. (f) Smartphone screening improvements after severe degradation. The improvement was computed by CausalFund - ERM.

gains were mixed but substantial in several settings (moderate hospital 0.065, severe smartphone 0.063, $p < 0.001$). Overall, the glaucoma screening results indicate robust benefit of CausalFund, especially under stronger smartphone-quality degradations.

For DR, CausalFund also outperformed ERM in most degradation levels, with mean improvements across all seven backbones concentrated on smartphone-domain screening (**Figure 4**). Mean AUC differences were positive in five of six settings (up to 0.027 for severe smartphone, $p < 0.05$), with slightly negative change only for severe hospital (-0.003). The composite Avg (Sens, Spec) metric was positive in five of six settings, with the largest gains for moderate smartphone (0.042, $p < 0.001$) and severe smartphone (0.030, $p < 0.001$), while

medRxiv preprint doi: https://doi.org/10.64898/2026.03.02.26347127; this version posted March 3, 2026. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

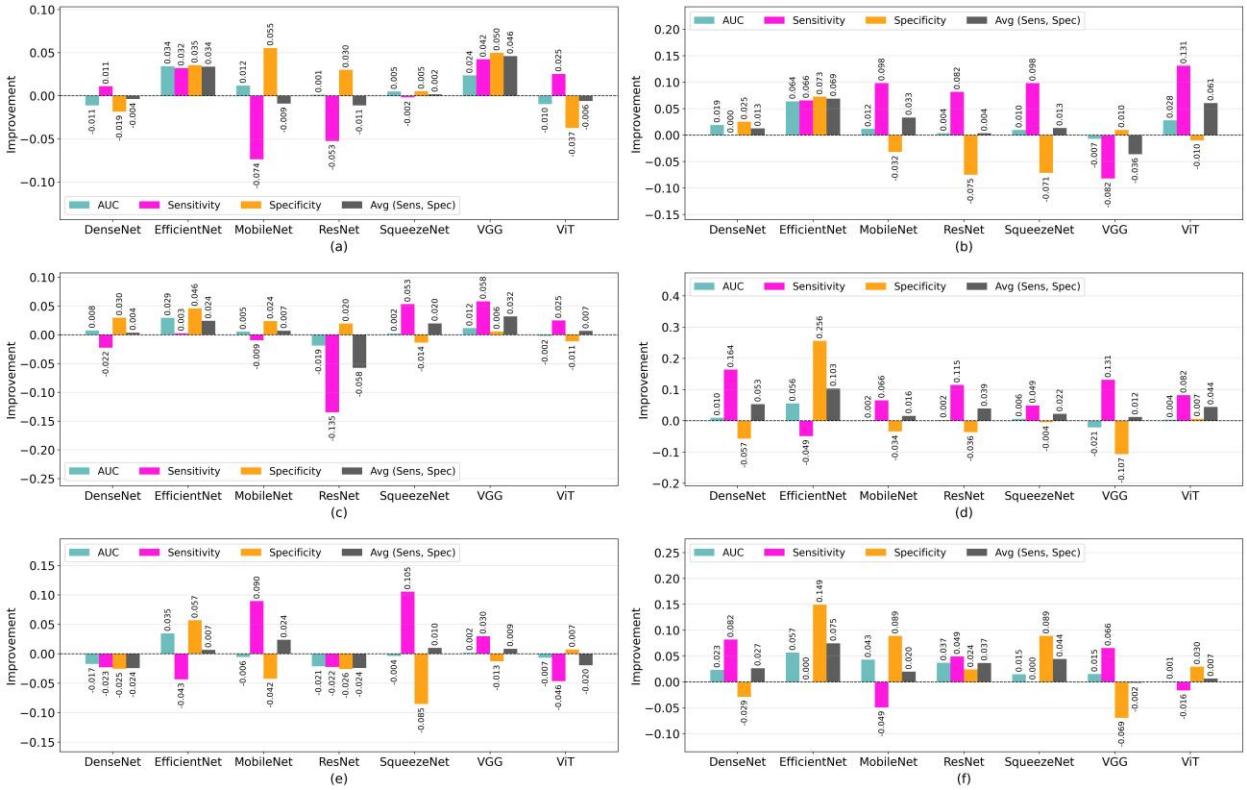


Figure 4: DR screening improvements of CausalFund over ERM after applying different levels (mild, moderate and severe) of degradation on smartphone images. (a) Hospital screening improvements after mild degradation. (b) Smartphone screening improvements after mild degradation. (c) Hospital screening improvements after moderate degradation. (d) Smartphone screening improvements after moderate degradation. (e) Hospital screening improvements after severe degradation. (f) Smartphone screening improvements after severe degradation. The improvement was computed by CausalFund - ERM.

severe hospital was near zero (-0.003). Sensitivity gains were strongest on smartphone screening (mild 0.056, moderate 0.080, severe 0.019, $p < 0.05$), and specificity gains were most pronounced for severe smartphone (0.041, $p < 0.001$) but mixed in some other settings. These results showed that CausalFund improved screening performance primarily under image-quality degradation in the smartphone domain.

For both glaucoma and DR, the image-quality degradation experiments demonstrated that CausalFund experienced much slower mean performance drops across all backbone models than ERM when smartphone image quality progressively turned worse from mild to severe levels (**Figure 5**). As an example, from mild to severe smartphone image quality variations (**Figure 6a**),

medRxiv preprint doi: https://doi.org/10.64898/2026.03.02.26347127; this version posted March 3, 2026. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

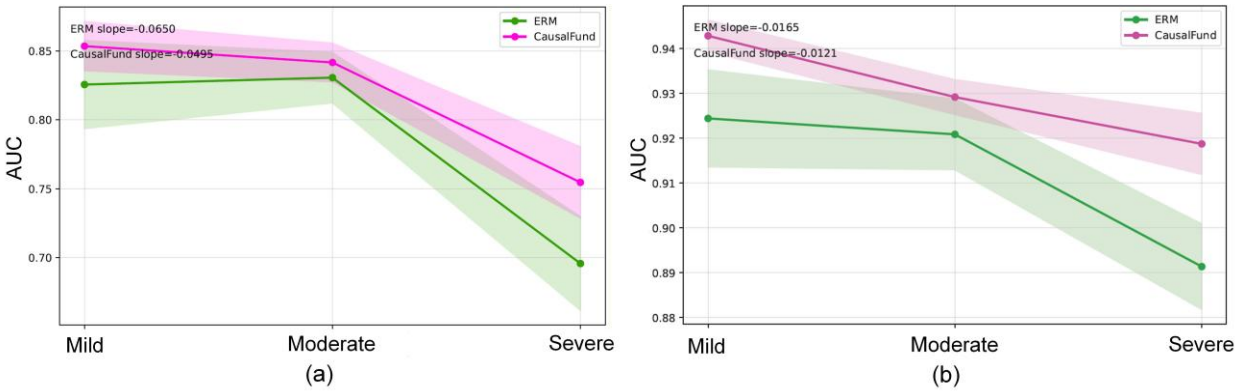


Figure 5: Performance degradation with increasing image-quality corruption in smartphone-based glaucoma and DR screening. (a) Mean AUC results of glaucoma screening. (c) Mean AUC results of DR screening. Slopes are linear fits to the three quality levels and shaded bands indicating SEM across all backbones for ERM and CausalFund at mild, moderate, and severe degradation levels. SEM: standard error of the mean.

the baseline ERM approach learned from varied regions caused by changing image qualities (**Figure 6b**), whereas CausalFund consistently focused on learning features at the optic nerve head area for glaucoma detection (**Figure 6c**).

Discussion

Recent years have witnessed a surge of AI models for automatically detecting retinal diseases such as glaucoma and DR.[20–22] Current works consistently demonstrate that AI-based screening has considerable potentials to reduce clinical care burden and cost to benefit broader populations, especially those from remote and low-resource areas.[6,23] However, the adoption of AI in real-world clinical environments remains hesitated and lagged behind. An undeniable factor lies in the fact that most state-of-the-art AI models are trained and evaluated using hospital-grade data from quality-guaranteed acquisition conditions, which generally suffer from poor generalizability to low-resourced settings where acquisition conditions and qualities are uncontrollable.[19,24] This study aims to mitigate this gap by training generalizable AI models for glaucoma and DR using low-cost fundus images acquired from portable devices and smartphones. The proposed CausalFund framework (**Figure 1c**) is a general framework for training domain-generalizable AI models and represents an effort to translate AI advances into realistic applications for cost-effective screening.

Our study showed that models trained in hospital domain suffer from significant performance degradation when applied in smartphone domain, which was especially prominent in poor imaging

medRxiv preprint doi: https://doi.org/10.64898/2026.03.02.26347127; this version posted March 3, 2026. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

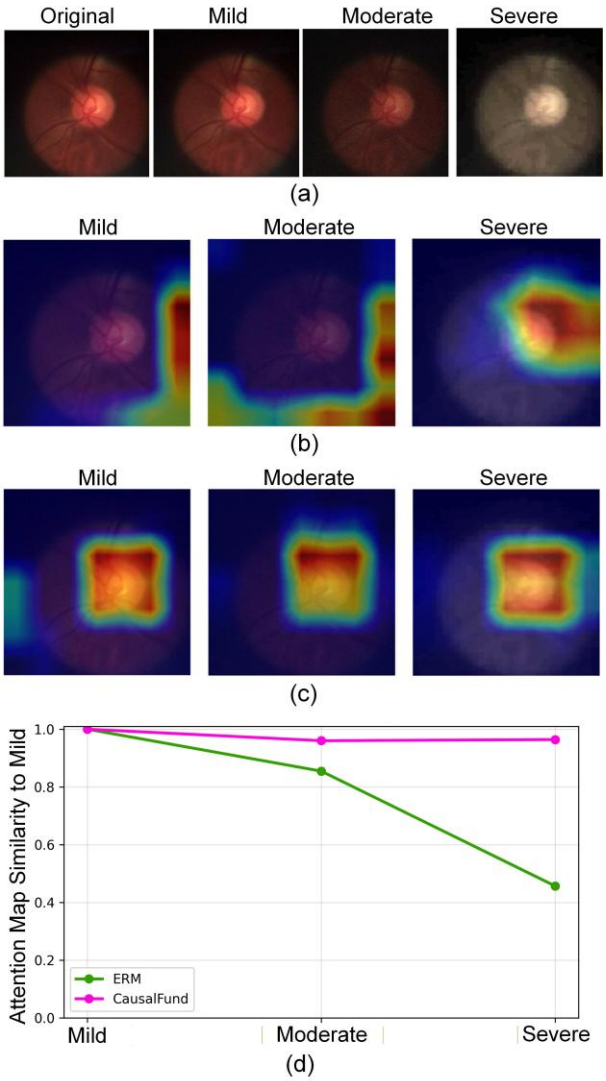


Figure 6: Grad-CAM comparison between ERM and CausalFund in smartphone-based glaucoma screening, and respective attention stabilities under progressive image-quality degradation - mild, moderate and severe. (a) The original smartphone image and images after applying mild, moderate and severe image-quality degradations. (b) Grad-CAM activation map of ERM (MobileNet backbone) under mild, moderate, and severe smartphone-quality degradations.

conditions. For example, under severe smartphone image-quality degradation, the averaged AUC gap between hospital and smartphone domains based on ERM across all seven backbones was 0.148 ($p < 0.001$), while this gap was significantly reduced to 0.100 ($p < 0.001$) with CausalFund which encouraged AI models to learn real causal features across domains for robust prediction. For both glaucoma and DR screening, CausalFund generally exhibited flatter degradation trends than ERM (**Figure 5**), and showed improved robustness to image-quality shifts. The advantage was attributed to the causal representation learning mechanism which enforced not to rely on

medRxiv preprint doi: https://doi.org/10.64898/2026.03.02.26347127; this version posted March 3, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

harmful domain-specific spurious factors, such as patient demographics and imaging lighting conditions, for decisions. Real causal features such as cup-to-disc ratios and neuroretinal rim thinning for reliably determining the presence of conditions should be consistent across domains and were explicitly learned in CausalFund. The case study showed that CausalFund was more robust to varying imaging conditions (e.g., spurious factors) and demonstrated stable learning capacity across domains (**Figure 6**), which is critical for low-resource settings where imaging conditions are often unpredictable and a generalizable model is desired.

Despite the advantages, our study has several limitations. First, evaluation was retrospective and based on curated public datasets, which may not fully represent real-world acquisition variability, device heterogeneity, and workflow constraints in routine screening. Second, image-quality shifts were modeled using predefined mild, moderate, severe augmentation pipelines. Although useful for controlled analysis, these synthetic degradations cannot capture all artifacts seen in practice (e.g., mixed blur, illumination instability, compression, and operator-dependent errors). Third, we focused on two tasks, glaucoma and DR, and a fixed set of backbones, so generalizability to other ophthalmic conditions, model families, and institutions remains to be established. Fourth, our interpretability analysis, Grad-CAM and attention-stability, is supportive but not a definitive causal explanation of model reasoning, and should be interpreted alongside task performance rather than as stand-alone evidence. Finally, we did not perform prospective clinical validation, calibration, fairness audits across demographic subgroups, or deployment-level analyses of latency and integration, which are important next steps before real-world adoption.

In conclusion, current deep learning models often use images from hospital-based devices, which can suffer from performance decline in low-resource settings due to uncontrolled, degraded image qualities. This work proposed a causality-inspired framework to enforce deep learning models to only learn causally generalizable retinal features across fundus imaging domains for robust screening. Empirical results and comparisons with traditional approaches demonstrated the effectiveness of CausalFund. The framework is extendable to other diseases to potentially achieve robust screening in low-resource imaging settings.

Methods

Four datasets were used to develop and evaluate deep learning models for glaucoma and DR screening. This study complied with the guidelines outlined in the Declaration of Helsinki. In light of the study's retrospective design, the requirement for informed consent was waived.

Dataset Description

medRxiv preprint doi: https://doi.org/10.64898/2026.03.02.26347127; this version posted March 3, 2026. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

For glaucoma, we used the dataset-provided annotations and formulated the task as binary classification (normal/healthy vs glaucoma). We enforced patient-level splitting to prevent identity leakage across train/validation/test. For the hospital glaucoma dataset, we parsed the provided metadata to derive patient identifiers and labels, removed mixed-label patients (i.e., patients with conflicting labels across their images), and then performed a stratified patient-level split (70%, 10%, 20%) into train/validation/test. For the BrG smartphone glaucoma dataset, we applied the same patient-level splitting strategy. For DR, labels follow the standard five-level grading scale (0-4: no DR, mild NPDR, moderate NPDR, severe NPDR, proliferative DR). We binarized DR severity into Non-DR (grades 0-2) versus DR (grades 3-4) and enforced this mapping consistently across the hospital and smartphone DR datasets. All images were loaded as RGB and resized to the model input size during training; standard data augmentation was applied to the training split only.

The Proposed CausalFund Framework

CausalFund is a model-agnostic and causality-inspired domain generalization framework designed to improve robustness when deploying fundus image screening models across acquisition domains, such as from hospital-grade cameras to portable/smartphone devices in this study. The main idea is to discourage reliance on domain-specific factors (e.g., patient demographics, illumination, contrast, blur) by explicitly training the model to be stable to controlled perturbations of its internal representation that simulate such spurious variation. Given an input fundus image, CausalFund contains following learning and oprimization processes (**Figure 1c**). First, the shared feature extractor, implemented as a deep learning backbone such as ResNet and ViT, learns image features into a latent causal representation Z. Then, an intervener, implemented as a three-layer MLP network, generates a label-conditioned (e.g., conditioned on the true label y or model-predicted label $\hat{y}$of the image) perturbation Δ intended to mimic domain shift in non-causal (spurious) factors while preserving the class semantics. This will result in an intervened latent representation $Z' = Z + \Delta$. Using the label (or predicted label) as the condition encourages the intervener to change nuisance characteristics without “crossing the decision boundary” into another disease class. Then, a shared classifier aims to predict class labels y and y' from causal representation Z and intervened representation, respectively. The model enforces predicted class labels $\hat{y}$and $\hat{y}'$ to be consistent, by which the model is optimized to use true causal representation Z for prediction without being affected by spurious features controlled by Δ. Take above processes together, CausalFund optimizes a combined loss that encourages both predictive performance and invariance to interventions:

medRxiv preprint doi: https://doi.org/10.64898/2026.03.02.26347127; this version posted March 3, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

$$L = L_{cls} + \lambda_{int}L_{int} + \lambda_{cons}L_{cons} + \lambda_{reg}L_{reg} + \lambda_{kl}L_{kl}, \quad (1)$$

where $L_{cls} = CE(\hat{y}, y)$ is classification loss on causal features Z. CE represents the cross-entropy (CE) loss measuring how well the model predictions match the true class labels.[25] $L_{int} = CE(\hat{y}', y)$ is intervened classification loss on intervened features to keep predictions correct under perturbation. $L_{cons} = \|\sigma(\hat{y}) - \sigma(\hat{y}')\|_2^2$ is consistency loss to encourage predictions from causal and intervened features to be consistent, where σ is sigmoid activation function. $L_{reg} = \left\|\|\Delta\|_2^2 - \sigma(\hat{y}')\right\|$ represents intervention regularization to constrain the magnitude of Δ with target scale b controlling typical intervention strength. $L_{kl} = D_{kl}(z \,||\, p(z|y))$ represents latent regularization, i.e., the Kullback–Leibler (KL) divergence D_{kl}, to discourage unstable latent representation learning, where $p(z|y)$ is a class-conditional prior. CausalFund conducts supervised training and optimizes with the Adam optimizer.[26]

To evaluate CausalFund, we treated hospital and smartphone images as separate environments. In the primary domain generalization setting, models were trained on the hospital training split, while smartphone data were reserved for validation and testing. We used a patient-level split of 70%, 10%, 20% for training, validation, testing, respectively. We evaluated CausalFund with both standard and lightweight deep learning backbones. Standard backbones included ResNet,[27] DenseNet,[28], EfficientNet,[29] and Vision Transformer (ViT),[30] while lightweight backbones included MobileNet,[31] and SqueezeNet.[32] For each backbone, we selected the best checkpoint across training epochs based on validation AUC, and reported test performance from the selected checkpoint.

Controlled Smartphone Image-Quality Degradation

To assess robustness under quality shift, we constructed three controlled degradation levels from the original smartphone images. The mild level applied Gaussian blur (radius = 1.2) and JPEG compression (quality = 50); the moderate level introduced downsample-upsample degradation (scale = 0.5), additive Gaussian noise (σ = 20), and reduced brightness (0.75) and contrast (0.85). The severe level combined motion blur (kernel size = 17), strong JPEG compression (quality = 15), and color distortion (hue shift = 0.07, saturation = 0.5). This progressively intensifying setup emulated realistic smartphone-specific artifacts, such as diminished sharpness, sensor noise, compression effects, and illumination or color instability.

Comparative Methods, Evaluation Metrics and Statistical Analysis

We compared CausalFund against empirical risk minimization (ERM) baselines using a consistent training and evaluation protocol across both glaucoma and diabetic retinopathy (DR) screening

medRxiv preprint doi: https://doi.org/10.64898/2026.03.02.26347127; this version posted March 3, 2026. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

tasks. For each task, we evaluated models built on several commonly used deep learning backbones for medical image analysis, including ResNet, DenseNet, EfficientNet, and Vision Transformer ViT, as well as lightweight architectures, MobileNet and SqueezeNet, suitable for resource-constrained deployment. Model performance was primarily assessed using the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity. For each backbone and method, the best checkpoint was selected based on validation AUC, and all reported test-set metrics were computed using the selected checkpoint. For cross-domain evaluation, we reported results separately on hospital and smartphone test sets to quantify the domain generalization gap. Statistical analysis focused on comparing AUC and threshold-based metrics between CausalFund and ERM under matched backbones and experimental settings. We used paired comparisons across backbones and experimental repetitions. We estimated uncertainty via bootstrap resampling on the test set. Statistical significance was evaluated using two-sided tests with a significance threshold of $p < 0.05$.

Parameter Settings

All models were trained using the same preprocessing and optimization pipeline to enable fair comparisons across methods and backbones. Images were loaded as RGB and resized to an input resolution of 224 × 224. Standard data augmentation (random crop, horizontal flip, rotation, and color jitter) was applied to the training split only, while validation and test images were evaluated using deterministic preprocessing (resize and normalization) without augmentation. We optimized model parameters using the Adam optimizer with weight decay disabled. Unless otherwise stated, models were trained for 40 epochs with a mini-batch size of 32 and a fixed random seed (seed = 0) to ensure reproducibility. CausalFund introduces additional parameters controlling the intervention module and regularization terms. For all backbones, the best checkpoint was selected based on validation AUC.

medRxiv preprint doi: https://doi.org/10.64898/2026.03.02.26347127; this version posted March 3, 2026. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

References

1. Ehrlich JR, Burke-Conte Z, Wittenborn JS, et al. Prevalence of Glaucoma Among US Adults in 2022. *JAMA Ophthalmol*. 2024;142(11):1046. doi:10.1001/jamaophthalmol.2024.3884
2. Chew EY. Foselutoclax (UBX1325) for Diabetic Macular Edema — A Potential Novel Therapy? *NEJM Evidence*. 2025;4(5). doi:10.1056/EVIDe2500004
3. Wong TY, Sun J, Kawasaki R, et al. Guidelines on Diabetic Eye Care. *Ophthalmology*. 2018;125(10):1608-1622. doi:10.1016/j.ophtha.2018.04.007
4. Bhargava S, Mason L, Okeke C. The Significance of Screening Family Members in Glaucoma: Opportunities and Challenges. *J Glaucoma*. 2024;33(8S):S40-S44. doi:10.1097/IJG.0000000000002400
5. Tang J, Liang Y, O'Neill C, Kee F, Jiang J, Congdon N. Cost-effectiveness and cost-utility of population-based glaucoma screening in China: a decision-analytic Markov model. *Lancet Glob Health*. 2019;7(7):e968-e978. doi:10.1016/S2214-109X(19)30201-3
6. Akune Y, Kawasaki R, Goto R, Tamura H, Hiratsuka Y, Yamada M. Cost-effectiveness of AI-based diabetic retinopathy screening in nationwide health checkups and diabetes management in Japan: A modeling study. *Diabetes Res Clin Pract*. 2025;221:112015. doi:10.1016/j.diabres.2025.112015
7. Huang JJ, Channa R, Wolf RM, et al. Autonomous artificial intelligence for diabetic eye disease increases access and health equity in underserved populations. *NPJ Digit Med*. 2024;7(1):196. doi:10.1038/s41746-024-01197-3
8. Sharma P, Takahashi N, Ninomiya T, et al. A hybrid multi model artificial intelligence approach for glaucoma screening using fundus images. *NPJ Digit Med*. 2025;8(1):130. doi:10.1038/s41746-025-01473-w
9. Hassan B, Raja H, Hassan T, et al. A comprehensive review of artificial intelligence models for screening major retinal diseases. *Artif Intell Rev*. 2024;57(5):111. doi:10.1007/s10462-024-10736-z
10. Ong Ly C, Unnikrishnan B, Tadic T, et al. Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data. *NPJ Digit Med*. 2024;7(1):124. doi:10.1038/s41746-024-01118-4
11. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med*. 2018;15(11):e1002683. doi:10.1371/journal.pmed.1002683
12. Wintergerst MWM, Jansen LG, Holz FG, Finger RP. A Novel Device for Smartphone-Based Fundus Imaging and Documentation in Clinical Practice: Comparative Image Analysis Study. *JMIR Mhealth Uhealth*. 2020;8(7):e17480. doi:10.2196/17480
13. Kim TN, Myers F, Reber C, et al. A Smartphone-Based Tool for Rapid, Portable, and Automated Wide-Field Retinal Imaging. *Transl Vis Sci Technol*. 2018;7(5):21. doi:10.1167/tvst.7.5.21
14. Geirhos R, Jacobsen JH, Michaelis C, et al. Shortcut learning in deep neural networks. *Nat Mach Intell*. 2020;2(11):665-673. doi:10.1038/s42256-020-00257-z
15. Chen RJ, Wang JJ, Williamson DFK, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat Biomed Eng*. 2023;7(6):719-742. doi:10.1038/s41551-023-01056-8

medRxiv preprint doi: https://doi.org/10.64898/2026.03.02.26347127; this version posted March 3, 2026. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

16. Yang Y, Zhang H, Gichoya JW, Katabi D, Ghassemi M. The limits of fair medical imaging AI in real-world generalization. *Nat Med*. 2024;30(10):2838-2848. doi:10.1038/s41591-024-03113-4
17. You C, Dai H, Min Y, Sekhon JS, Joshi S, Duncan JS. Uncovering memorization effect in the presence of spurious correlations. *Nat Commun*. 2025;16(1):5424. doi:10.1038/s41467-025-61531-5
18. Scholkopf B, Locatello F, Bauer S, et al. Toward Causal Representation Learning. *Proceedings of the IEEE*. 2021;109(5):612-634. doi:10.1109/JPROC.2021.3058954
19. Lv F, Liang J, Li S, et al. Causality Inspired Representation Learning for Domain Generalization. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2022:8036-8046. doi:10.1109/CVPR52688.2022.00788
20. Ashtari-Majlan M, Dehshibi MM, Masip D. Glaucoma diagnosis in the era of deep learning: A survey. *Expert Syst Appl*. 2024;256:124888. doi:10.1016/j.eswa.2024.124888
21. Coan LJ, Williams BM, Krishna Adithya V, et al. Automatic detection of glaucoma via fundus imaging and artificial intelligence: A review. *Surv Ophthalmol*. 2023;68(1):17-41. doi:10.1016/j.survophthal.2022.08.005
22. Vaughan N. Review of smartphone funduscopy for diabetic retinopathy screening. *Surv Ophthalmol*. 2024;69(2):279-286. doi:10.1016/j.survophthal.2023.10.006
23. Xiao X, Xue L, Ye L, Li H, He Y. Health care cost and benefits of artificial intelligence-assisted population-based glaucoma screening for the elderly in remote areas of China: a cost-offset analysis. *BMC Public Health*. 2021;21(1):1065. doi:10.1186/s12889-021-11097-w
24. Luo Yan, Shi Min, Tian Yu, Tobias Elze, Mengyu Wang. Harvard Glaucoma Detection and Progression: A Multimodal Multitask Dataset and Generalization-Reinforced Semi-Supervised Learning. In: IEEE; 2023:20471-20482.
25. Martinez M, Stiefelhagen R. Taming the Cross Entropy Loss. In: 2019:628-637. doi:10.1007/978-3-030-12939-2_43
26. Zhang Z. Improved Adam Optimizer for Deep Neural Networks. In: *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*. IEEE; 2018:1-2. doi:10.1109/IWQoS.2018.8624183
27. He K, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep residual learning for image recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2016:770-778.
28. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2017:2261-2269. doi:10.1109/CVPR.2017.243
29. Tan M, Le Q V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Published online May 28, 2019.
30. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Published online October 22, 2020.
31. Howard AG. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:170404861*. Published online 2017.
32. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv preprint arXiv:160207360*. Published online 2016.

medRxiv preprint doi: https://doi.org/10.64898/2026.03.02.26347127; this version posted March 3, 2026. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Table 1. Dataset characteristics.

Dataset	Domain	Items	Statistics
Glaucoma screening	Hospital	Data type	Color fundus photographs
		Label type	Binary (normal/healthy vs glaucoma)
		Class distribution	0/healthy 7,549 (61.3%), 1/glaucoma 4,767 (38.7%)
		Train / Val / Test (images)	8,188 / 1,018 / 3,110 (total 12,316)
Glaucoma screening	Smartphone	Data type	Portable/smartphone fundus photographs
		Label type	Binary (normal/healthy vs glaucoma)
		Class distribution	0/healthy 1,000 (50.0%), 1/glaucoma 1,000 (50.0%)
		Train / Val / Test (images)	1,400 / 200 / 400 (total 2,000)
DR screening	Hospital	Data type	Color fundus photographs
		Label type	Binary DR (Non-DR: grades 0-2 vs DR: grades 3-4)
		Class distribution	0/Non-DR 28,173 (62.2%), 1/DR 17,154 (37.8%)
		Train / Val / Test (images)	29,840 / 6,796 / 8,691 (total 45,327)
DR screening	Smartphone	Data type	Portable/smartphone fundus photographs (mBRSET)
		Label type	Binary DR (Non-DR: grades 0-2 vs DR: grades 3-4)
		Class distribution	0/Non-DR 4,590 (94.0%), 1/DR 294 (6.0%)
		Train / Val / Test (images)	3,420 / 493 / 971 (total 4,884)