Bridging the Ophthalmic Screening Gap: Evaluating Diffusion-Based and Classical Preprocessing for Low-Resource Eye Disease Classification

Hayoon Kim
Korean Minjok Leadership Academy

ABSTRACT

Automated retinal disease screening in low-resource settings is constrained by the image quality limitations of low-cost fundus cameras, yet the impact of realistic degradation on multi-class classification performance remains poorly characterized. This study investigates whether DDPM-based restoration can stabilize retinal disease classification from degraded fundus photographs. Using a four-class fundus dataset (4,217 images: normal, cataract, diabetic retinopathy, glaucoma), we trained an EfficientNet-B0 classifier under four conditions — high-quality (HQ), synthetically degraded low-quality (LQ), LQ+CLAHE, and LQ+DDPM — evaluated via three-fold stratified cross-validation. Synthetic degradation reduced macro-F1 by 6.0 percentage points and macro-AUC by 1.8 points relative to HQ. Among preprocessing strategies, DDPM-based restoration achieved macro-F1 of 0.8807 and macro-AUC of 0.9780, outperforming both the LQ baseline (F1 0.8587) and CLAHE (F1 0.8604), while reducing cross-fold F1 standard deviation from 0.0174 to 0.0056 — a three-fold improvement in generalization stability. Qualitative analysis further confirmed that DDPM enhances vessel visibility and optic disc clarity without the noise overamplification occasionally introduced by CLAHE. These findings position DDPM-based preprocessing as a compelling strategy for robust retinal disease classification in resource-constrained screening environments.

KEYWORDS

retinal fundus imaging, diffusion-based image restoration, ophthalmic disease classification, low-resource screening, generalization stability

1. INTRODUCTION

Blindness and severe visual impairment represent a major global public health burden, with the World Health Organization estimating that at least 2.2 billion people worldwide suffer from vision impairment, a disproportionate share residing in low- and middle-income countries where access to specialized ophthalmic care remains critically limited. [1] Diabetic retinopathy (DR), glaucoma, and cataract are among the leading causes of avoidable blindness, yet each can be substantially prevented or delayed through early detection. [2] Despite this, a large proportion of cases are diagnosed only after irreversible visual field loss has occurred — a failure that systematic screening could prevent.

Fundus photography-based screening has been established as the cornerstone strategy for early detection of these conditions, enabling non-invasive visualization of the retina, optic disc, macula, and vascular architecture within a single imaging session. [3] However, high-end tabletop fundus cameras — typically costing between $15,000 and $50,000 — remain inaccessible in most low-income and rural settings, rendering population-level screening programs difficult to implement [4] and giving rise to a persistent health equity gap in global ophthalmic care.

Among the retinal conditions amenable to fundus-based screening, each disease presents distinct and diagnostically characteristic visual signatures. Diabetic retinopathy manifests as microaneurysms — small, dot-like hemorrhages arising from weakened capillary walls — alongside intraretinal hemorrhages, hard exudates, and, in advanced stages, neovascularization along the optic disc and retinal periphery. [5] Glaucoma is characterized by progressive excavation of the optic disc, reflected in an increased cup-to-disc ratio and thinning of the neuroretinal rim, which corresponds to irreversible loss of retinal ganglion cells. [6] Cataract, while primarily a lens opacity rather than a retinal pathology, manifests in fundus photographs as a diffuse reduction in image clarity and contrast caused by light scattering through the clouded lens. [7] Normal fundus images, by contrast, display a well-defined optic disc with a small physiological cup, uniform vessel caliber, and a clearly visible foveal reflex. [3] These disease-specific visual features — spanning fine microstructural changes at the pixel scale in DR to global structural alterations in glaucoma — underscore why image quality is not a peripheral concern but a direct determinant of diagnostic reliability in automated screening systems. [8]

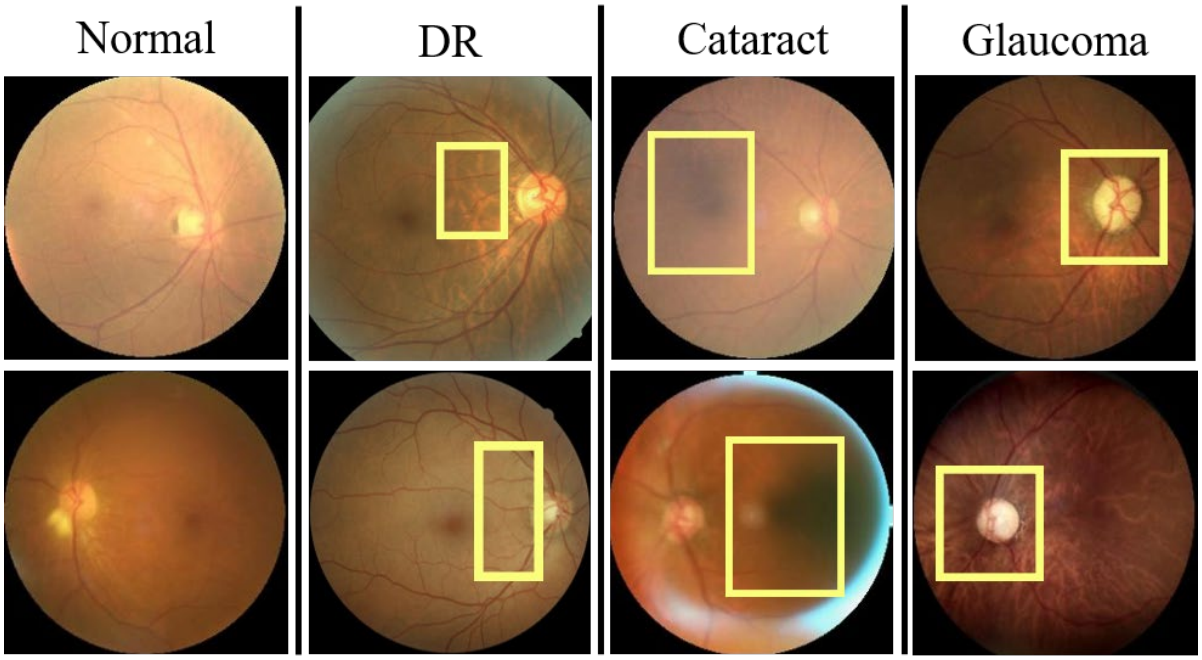

Fig 1. Representative fundus images across four diagnostic categories
Red boxes indicate class-specific pathological features: in Diabetic Retinopathy, hard exudates and peripapillary lesions near the optic disc; in Cataract, diffuse image haziness and loss of vascular definition due to lens opacity; in Glaucoma, enlarged optic disc cup reflecting increased cup-to-disc ratio. Normal retinas show no annotated abnormalities.
**Source: Generated by the authors using Python (Matplotlib) on KISTI supercomputing infrastructure; figures compiled and edited in Microsoft PowerPoint.*

Low-cost smartphone-attachable and portable fundus cameras have emerged as a promising alternative, priced as low as a few hundred dollars and deployable at the community level. [9] However, images captured by such devices frequently exhibit quality degradations including resolution loss, blur, sensor noise, JPEG compression artifacts, and uneven illumination. [10] When deep learning-based classification models are applied to such low-quality (LQ) images, they consistently fail to reproduce the accuracy reported on high-quality (HQ) datasets. [11] Diseases relying on fine structural cues are especially vulnerable: DR requires detection of microaneurysms at the pixel scale, while glaucoma depends on subtle optic disc morphology — both of which are readily obscured by moderate blur or resolution loss. [12]

Prior deep learning studies have achieved approximately 95% classification accuracy across DR, glaucoma, and cataract using EfficientNet-based architectures, but largely assume idealized acquisition conditions. [13] Systematic evaluation of disease-specific failure patterns under realistic camera degradation, and rigorous comparison of preprocessing strategies to close the resulting performance gap, remain insufficient. Meanwhile, diffusion models have recently demonstrated strong performance in low-quality image restoration, with Denoising Diffusion Probabilistic Models (DDPMs) capable of preserving vascular structures, optic disc morphology, and lesion patterns through iterative generative denoising — operating fundamentally differently from classical methods such as CLAHE, which merely redistributes existing pixel intensities rather than reconstructing lost image content. [14] Yet whether diffusion-restored images actually improve multi-class ophthalmic screening performance, and under what conditions, remains poorly characterized.

This study addresses these gaps by leveraging a dataset of 4,217 fundus images spanning four categories (Normal: 1,074; Cataract: 1,038; Diabetic Retinopathy: 1,098; Glaucoma: 1,007) to conduct a comparative analysis between HQ images and synthetically degraded LQ images simulating low-cost camera conditions. [15] We investigate three research questions.

First (RQ1), to what extent does LQ degradation reduce the screening performance of a four-class ophthalmic classifier, and how does this vary across disease types and degradation levels?

Second (RQ2), how effectively does DDPM-based restoration preprocessing (LQ+Diff) reduce the performance gap between LQ and HQ conditions, and what additional benefit does it provide over classical CLAHE-based enhancement (LQ+CLAHE)?

Third (RQ3), under what disease and degradation conditions is diffusion restoration particularly beneficial, and how does it compare to CLAHE in terms of generalization stability and consistency across data splits?

By answering these three questions, this study aims to provide a disease-stratified and condition-specific guide to when and how diffusion-based preprocessing can be safely and effectively introduced for ophthalmic screening in low-resource settings. In doing so, we offer a practical framework for bridging the image quality gap that currently limits equitable access to AI-assisted eye care across the globe.

2. RELATED WORK

Deep learning has rapidly become the dominant approach for ophthalmic disease classification from fundus images, with CNNs and transformer-based models now achieving high accuracy for detecting diabetic retinopathy, glaucoma, and cataract from color fundus photographs (Li et al). [16] EfficientNet-based classifiers have been reported to achieve over 94–95% accuracy for multi-class ophthalmic screening under high-quality imaging conditions. [17] However, these works largely assume idealized acquisition conditions and do not examine performance under the degradations characteristic of low-cost portable devices — the very scenario most relevant to equitable screening deployment. [11]

Image quality and preprocessing have therefore emerged as critical topics in fundus-based AI. CLAHE has been widely adopted as a preprocessing step, with Huang et al., Patel et al., and Singh et al. demonstrating consistent improvements in local contrast and CNN-based classification performance across DR and glaucoma datasets. [18] Beyond classical methods, Zhang et al. showed that more structured preprocessing pipelines — combining color-space transformations or multi-stage enhancement — can further improve downstream classification and segmentation. [19] Collectively, these studies confirm that preprocessing meaningfully influences AI screening performance, while also highlighting that classical methods have diminishing returns when degradations are severe or multi-factorial. [19]

More recently, diffusion models have demonstrated strong performance in medical image restoration. [20] Park et al. showed that diffusion-based reconstruction outperforms GAN-based approaches across multiple imaging modalities in PSNR, SSIM, and perceptual quality, particularly under complex or heterogeneous degradations. [21] In fundus imaging specifically, Liu et al. introduced a unified enhancement framework for low-quality handheld fundus images, Go et al. proposed FD3 — a diffusion bridge method co-designed with ophthalmologists that consistently outperforms CLAHE across challenging cases including cataract and small-pupil imaging — and Yang et al. presented a retinal latent diffusion framework that enables efficient high-resolution enhancement while preserving fine vascular and lesion structures across heterogeneous datasets. [22]

Despite these advances, critical gaps remain. Most diffusion-based fundus enhancement studies evaluate performance using image-level metrics such as PSNR and SSIM, or intermediate tasks like vessel segmentation, rather than directly assessing impact on multi-class disease classification. [23] Few works simulate low-cost camera conditions through controlled degradation pipelines, and direct comparisons between diffusion-based and classical preprocessing under matched classification protocols remain scarce. [23] To the best of our knowledge, no prior study has systematically compared DDPM-based and CLAHE-based preprocessing for a four-class ophthalmic classifier under explicitly simulated low-quality conditions. This work addresses these gaps through a controlled pipeline evaluated through a screening-oriented lens that prioritizes classification performance, generalization stability, and safety in low-resource settings.

3. METHODS

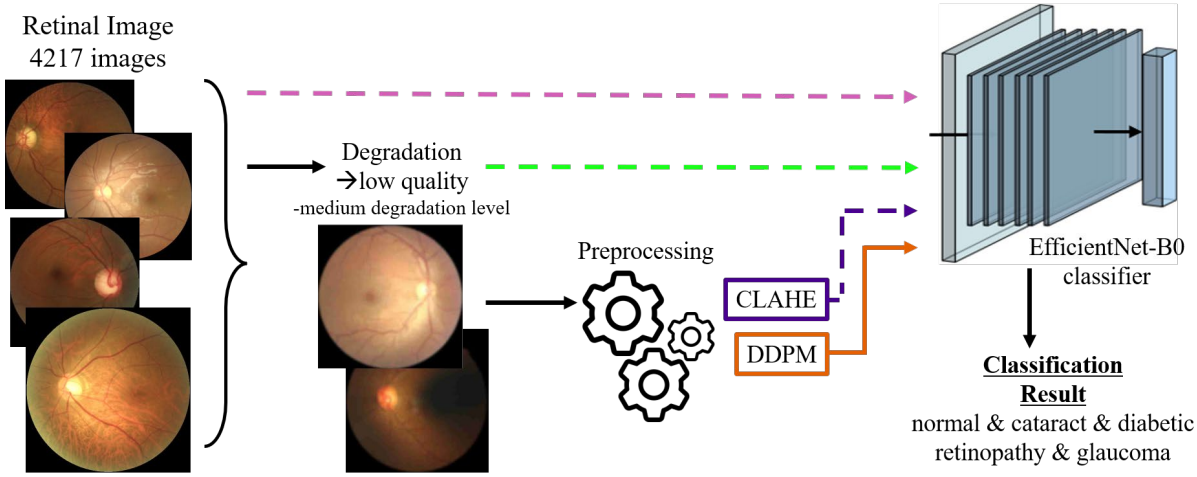

Fig 2. Overview of the proposed preprocessing pipeline for retinal disease classification
**Source: Compiled by the authors; figures created in Microsoft PowerPoint.*

3.1 Dataset and Problem Setup

We base our experiments on the Eye Diseases Classification dataset publicly available on Kaggle, which contains 4,217 color fundus images labeled into four diagnostic categories: Normal (1,074 images), Cataract (1,038), Diabetic Retinopathy (1,098), and Glaucoma (1,007). [15] The dataset originates from multiple clinical sources and exhibits variability in field of view, illumination, and retinal pigmentation, providing a realistic testbed for multi-class ophthalmic screening. Prior work using this dataset has demonstrated that EfficientNet-based architectures achieve approximately 92–95% accuracy under high-quality imaging conditions [13], establishing a strong baseline for our quality-degradation analysis.

Setting	Details
Dataset	4,217 fundus images, 4 classes
Validation	3-fold stratified cross-validation
Degradation	downsample + Gaussian blur + JPEG compression
Augmentation	flip, rotation(±15°), color jitter, affine transform
Classifier	EfficientNet-B0, ImageNet pretrained
DDPM	Unet, 128 X 128, unconditional generation

Table 1. Dataset and experimental setup summary
**Source: Compiled by the authors based on experimental results; tables created in Microsoft PowerPoint.*

We adopt three-fold cross-validation stratified by class, with two folds used for training and one held out as the test set. For each split, the training portion is further divided into training and validation subsets with an 85:15 ratio. The validation set is used for early stopping and model selection. Class-weighted loss compensates for minor residual imbalances without oversampling. Standard data augmentation — including random flips, rotations (±15°), color jitter, and affine transformations — is applied consistently across all preprocessing conditions. All images are resized to 224 × 224 pixels.

3.2 Low-Quality Simulation Pipeline

To model image degradation from low-cost portable fundus cameras, we design a synthetic LQ simulation pipeline applying three sequential transformations to each HQ image: resolution downsampling followed by bilinear upsampling to 224 × 224, isotropic Gaussian blurring (std σ), and JPEG re-encoding at quality factor Q.

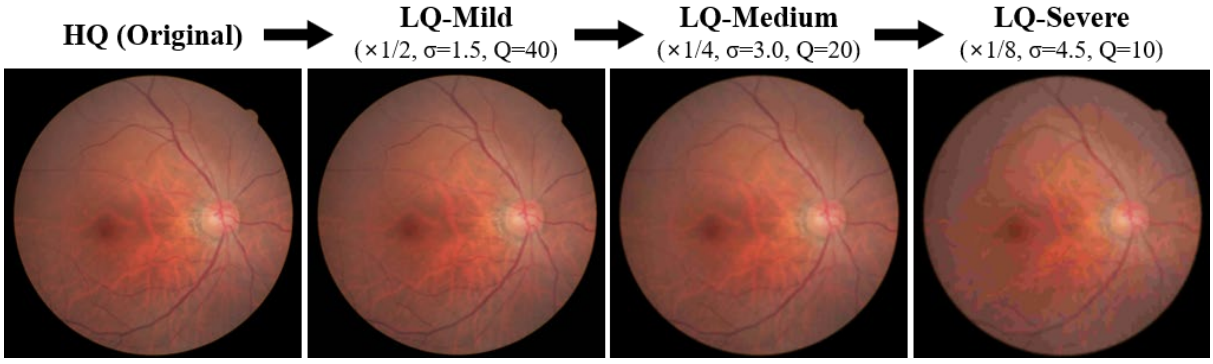

Fig 3. Synthetic low-quality simulation pipeline across three degradation levels
**Source: Generated by the authors using Python (Matplotlib) on KISTI supercomputing infrastructure; figures compiled and edited in Microsoft PowerPoint.*

Three degradation levels are defined as follows:

Level	Downsample factor	Gaussian σ	JPEG Q
Mild	1/2	1.5	40
Medium	1/4	3.0	20
Severe	1/8	4.5	10

Table 2. Degradation parameters for each severity level
**Source: Compiled by the authors based on experimental results; tables created in Microsoft PowerPoint.*

These values were determined through preliminary visual inspection. The same parameters are applied across all four disease classes to enable disease-stratified analysis. In our primary analysis, we focus on the medium degradation level as the main experimental condition. The simulation is implemented offline prior to all training and evaluation, ensuring reproducibility and preventing data leakage.

3.3 Preprocessing Methods

We compare three preprocessing pipelines applied to LQ images under identical classification architectures and training schedules.

Baseline. Images are resized to 224 × 224 and normalized using ImageNet statistics, with no enhancement applied.

CLAHE-based enhancement. CLAHE is applied in the LAB color space on the luminance channel with clip limit 2.0 and tile grid size 8 × 8, following widely adopted settings in retinal image processing. [18] Enhanced images are converted back to RGB and normalized identically to the baseline.

Diffusion-based restoration. We train a DDPM (Ho et al., 2020) from scratch on our 4,217-image fundus dataset using the U-Net implementation from Rampas (2022), at 128 × 128 resolution for 30 epochs (AdamW, lr = 3 × 10^{-4}, batch size 4). [24] Training directly on our domain-specific dataset ensures the learned image prior reflects the distributional properties of ophthalmic fundus photographs. For restoration, each LQ image is downscaled to 128 × 128, partially noised to timestep t = 100, and denoised through the reverse process. The output is upsampled to 224 × 224 via bicubic interpolation and blended with the original LQ image (α = 0.7) to preserve global anatomical structure while suppressing degradation artifacts. We denote these images as LQ+Diff.

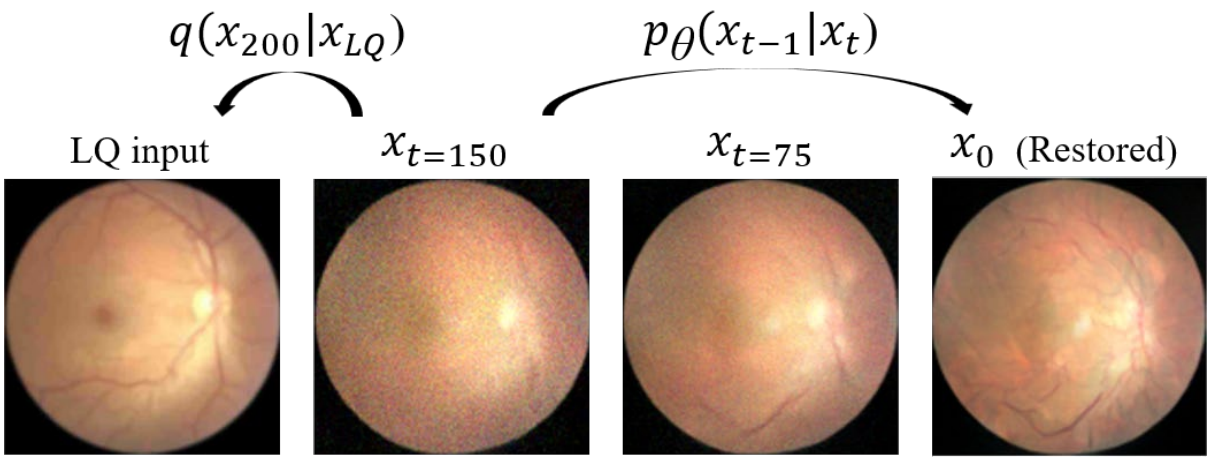

Fig 4. DDPM Denoising Trajectory on a Representative Fundus Image
**Source: Generated by the authors using Python (Matplotlib) on KISTI supercomputing infrastructure; figures compiled and edited in Microsoft PowerPoint.*

3.4 Classification Model and Evaluation

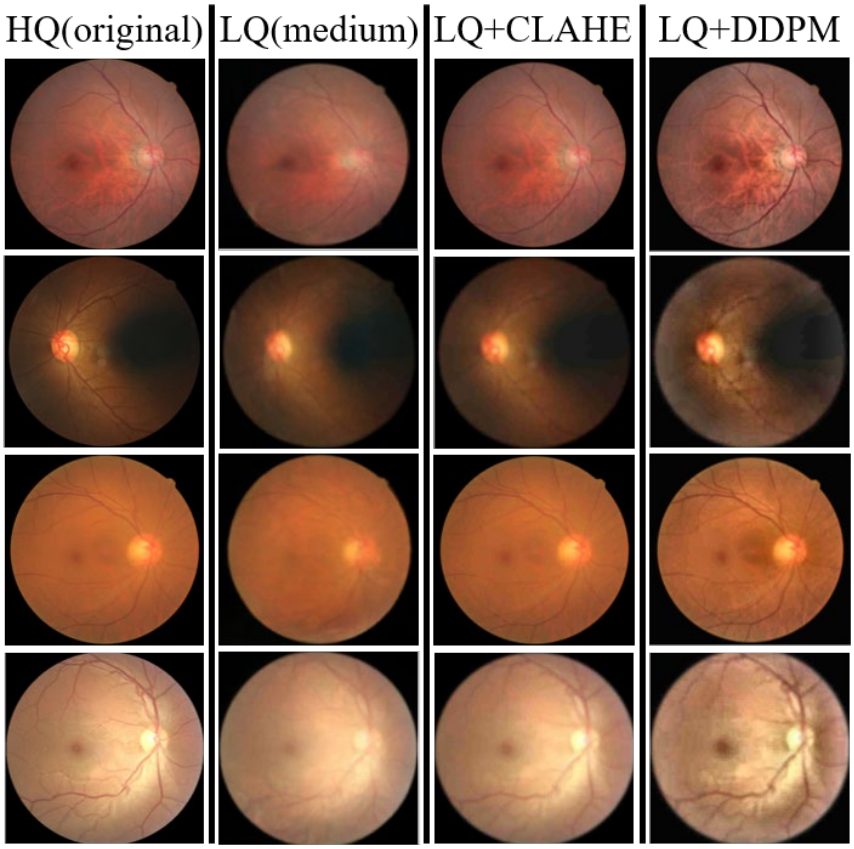

Fig 5. Representative fundus images under each preprocessing condition across four disease classes
**Source: Generated by the authors using Python (Matplotlib) on KISTI supercomputing infrastructure; figures compiled and edited in Microsoft PowerPoint.*

We employ EfficientNet-B0 pretrained on ImageNet as the base classifier, replacing the final layer with a four-unit output corresponding to Normal, Cataract, Glaucoma, and Diabetic Retinopathy. A separate model is trained per preprocessing condition per fold. Training uses AdamW (lr = 1×10^{-4}, weight decay = 1×10^{-4}) with cosine annealing, batch size 64, and early stopping on validation Macro-F1 (patience = 5, max 30 epochs). Class-weighted CrossEntropyLoss compensates for minor label imbalances. We report Accuracy, Macro-F1, and Macro-AUC averaged across three folds with mean and standard deviation.

RESULTS

	Accuracy	Macro-F1	Macro-AUC
High Quality (HQ)	0.9151 ± 0.0055	0.9184 ± 0.0048	0.9882 ± 0.0020
Low Quality (LQ)	0.8601 ± 0.0079	0.8587 ± 0.0073	0.9702 ± 0.0020
LQ + CLAHE	0.8572 ± 0.0173	0.8604 ± 0.0174	0.9706 ± 0.0123
LQ + DDPM	0.8793 ± 0.0073	0.8807 ± 0.0056	0.9780 ± 0.0008

Table 3. Three-fold cross-validation results across preprocessing conditions
**Source: Compiled by the authors based on experimental results; tables created in Microsoft PowerPoint.*

4.1 Effect of Image Degradation on Classification Performance

To quantify the impact of synthetic image degradation on automated retinal disease classification, we first compared model performance between high-quality (HQ) fundus images and their low-quality (LQ) counterparts generated through the degradation pipeline described in Section 3.2, without applying any preprocessing enhancement. Across three-fold cross-validation using EfficientNet-B0, the HQ condition achieved an accuracy of 0.9151 ± 0.0055, a macro-F1 score of 0.9184 ± 0.0048, and a macro-AUC of 0.9882 ± 0.0020. These results are consistent with prior benchmarks on the same dataset, where EfficientNet-based classifiers have been reported to achieve validation accuracies of 92–95% under clean imaging conditions, confirming that our experimental setup yields competitive baseline performance. [13]

In contrast, training and evaluation on the medium-degradation LQ images resulted in consistently lower performance across all three metrics: accuracy of 0.8601 ± 0.0079, F1-score of 0.8587 ± 0.0073, and AUC of 0.9702 ± 0.0020. The absolute drop of approximately 6.0 percentage points in F1 and 1.8 points in AUC relative to the HQ condition is both statistically consistent across folds and practically meaningful in a screening context, indicating that the combined effect of resolution reduction, Gaussian blur, and JPEG compression introduces sufficient information loss to meaningfully impair classification performance.

From a disease-stratified perspective, this performance gap aligns with the known sensitivity of different ophthalmic conditions to image quality degradation. Diabetic retinopathy, which relies on the detection of microaneurysms and fine intraretinal hemorrhages, is particularly susceptible to resolution loss and blur. Similarly, glaucoma diagnosis depends on subtle morphological changes around the optic disc margin that are readily obscured by moderate degradation. In contrast, cataract, which already manifests as diffuse image haze, may be relatively less sensitive to additional synthetic artifacts.

It is also notable that despite the performance drop, an F1 of 0.8587 on LQ images indicates that a substantial portion of clinically meaningful retinal structure remains recoverable under moderate degradation. This suggests that preprocessing enhancement can meaningfully close the gap — as explored in Section 4.2 — though the degree of recovery will depend on the method's ability to restore diagnostically relevant features rather than simply improving visual appearance.

Taken together, the HQ versus LQ comparison provides a clear quantitative answer to RQ1: moderate synthetic degradation reduces four-class ophthalmic classification performance by approximately 6 percentage points in F1 and 1.8 points in AUC, with low inter-fold variability confirming that this effect is robust and reproducible across different data splits.

4.2 Preprocessing Comparison: CLAHE vs Diffusion-Based Restoration

To investigate whether preprocessing can mitigate the performance drop induced by image degradation, we applied two enhancement strategies to the LQ images prior to classification: contrast-limited adaptive histogram equalization (CLAHE) and diffusion-based restoration using a denoising diffusion probabilistic model (DDPM) trained on the fundus dataset. Both pipelines were evaluated under identical classification model architectures, training schedules, and cross-validation protocols to ensure that any observed differences in performance are attributable solely to the preprocessing method.

When LQ images were preprocessed with CLAHE, the classifier achieved an accuracy of 0.8572 ± 0.0173, an F1-score of 0.8604 ± 0.0174, and an AUC of 0.9706 ± 0.0123. Diffusion-based restoration yielded an accuracy of 0.8793 ± 0.0073, an F1-score of 0.8807 ± 0.0056, and an AUC of 0.9780 ± 0.0008. Unlike CLAHE, which produced performance broadly comparable to the unenhanced LQ baseline, DDPM-based restoration achieved

a meaningful improvement across all three metrics — recovering approximately 2.2 percentage points in F1 and 0.8 points in AUC relative to LQ. The diffusion-based approach also outperformed CLAHE by 2.0 percentage points in F1 (0.8807 versus 0.8604), suggesting that generative restoration provides a tangible and consistent advantage over conventional contrast enhancement under these experimental conditions.

The most analytically significant difference between the two methods, however, emerges not only in mean performance but also in cross-fold variability. CLAHE exhibited a standard deviation of 0.0174 in F1 and 0.0123 in AUC, indicating notable instability across the three cross-validation splits. In contrast, DDPM-based restoration produced a standard deviation of only 0.0056 in F1 and 0.0008 in AUC — a reduction in F1 variance by more than a factor of three relative to CLAHE. This difference is noteworthy because high cross-fold variance implies that a method's effectiveness depends strongly on the particular distribution of training and validation data, raising concerns about generalizability to unseen patient populations or imaging conditions. The DDPM pipeline's markedly lower variance suggests that generative restoration produces more consistent image transformations, resulting in a downstream classifier that generalizes more reliably across different data partitions.

It is worth contextualizing these findings within the broader landscape of diffusion-based fundus enhancement. Prior work has demonstrated that diffusion models can achieve superior PSNR and SSIM on synthetically degraded fundus images, and ophthalmologist-rated evaluations have consistently favored diffusion-enhanced images over CLAHE-processed counterparts in terms of vessel visibility and diagnostic usability. [22] Our classifier-based evaluation corroborates and extends these trends: DDPM-restored images not only support superior classification performance compared to CLAHE, but do so with substantially lower variance, reinforcing the practical case for generative preprocessing in heterogeneous screening environments. That said, neither method fully recovered the HQ baseline performance, consistent with the known limitations of unconditional DDPM under resource-constrained training conditions.

A critical factor shaping our results is the design of the DDPM pipeline. Our model was trained at 128 × 128 resolution on 4,217 fundus images — substantially fewer data and lower resolution than the large-scale paired datasets used in state-of-the-art diffusion-based fundus enhancement studies. [22] The use of an unconditional DDPM, which learns the marginal distribution of clean fundus images without explicit conditioning on LQ inputs, means that restoration operates through partial noising and denoising rather than a directed LQ-to-HQ mapping. Under these constraints, the DDPM functions as a structural regularizer that draws restored images toward the learned fundus image prior — and even in this configuration, it achieves meaningful classification gains over both the LQ baseline and CLAHE. This suggests that the observed improvements reflect a genuine capacity of diffusion-based priors to recover diagnostically relevant image structure, and that further gains are plausible with higher-resolution training, larger datasets, or conditional architectures that explicitly model the LQ-to-HQ transformation.

Taken together, these findings answer RQ2 by demonstrating that under resource-constrained conditions, diffusion-based restoration delivers meaningfully superior average classification performance compared to CLAHE, while also exhibiting substantially lower fold-to-fold variance. The advantages of DDPM in this setting span both accuracy and generalization stability, suggesting a practical and compelling role for diffusion preprocessing in low-resource ophthalmic screening deployments — particularly where robustness across diverse imaging conditions is essential.

4.3 Qualitative Analysis

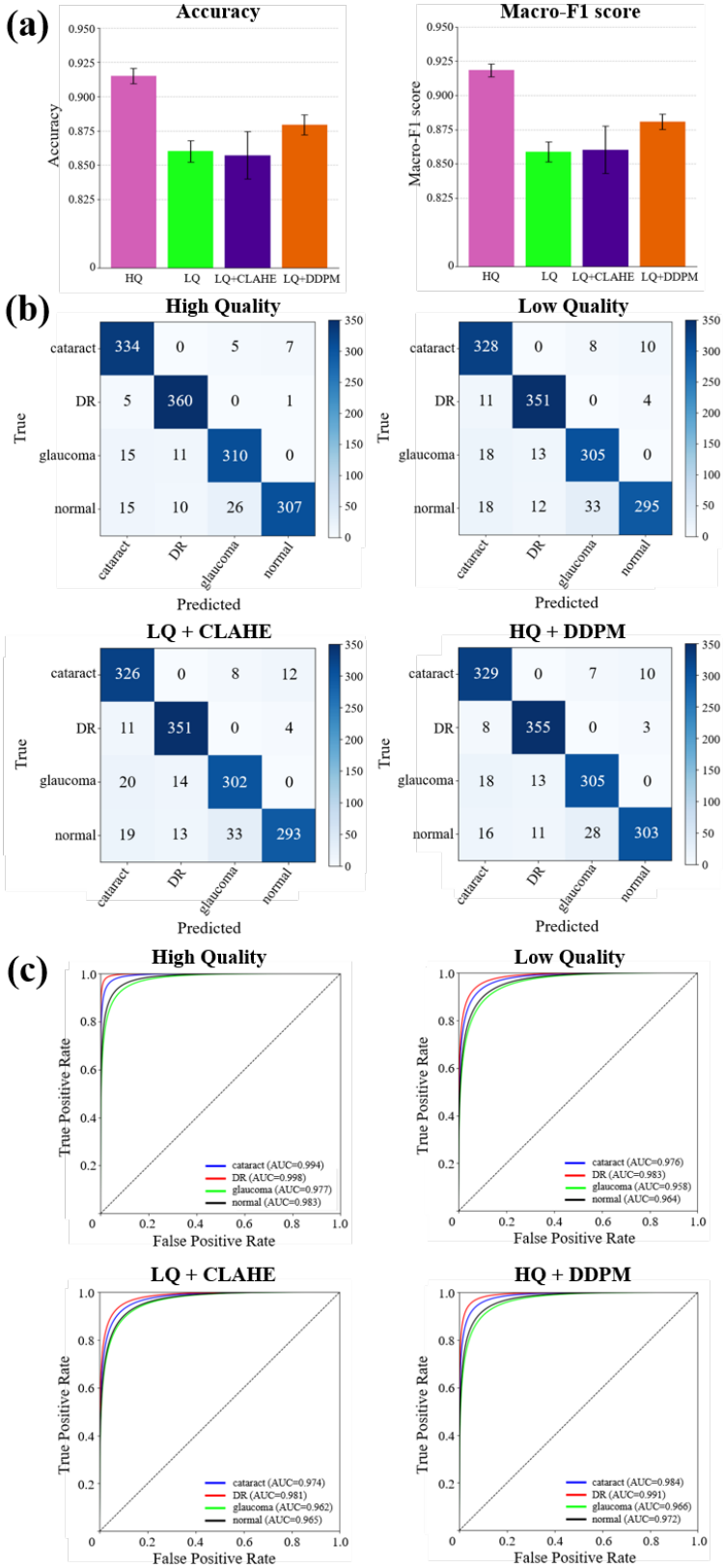

Fig 6. Classification Performance Across Preprocessing Conditions (Mean ± Std, 3-Fold Stratified Cross-Validation)

(a) Overall accuracy and macro-F1 score across preprocessing conditions (b) Average confusion matrices across folds for each preprocessing condition (c) Per-class ROC curves and AUC scores for each preprocessing condition

**Source: Generated by the authors using Python (Matplotlib) on KISTI supercomputing infrastructure; figures compiled and edited in Microsoft PowerPoint.*

Qualitative inspection of representative fundus images provides additional context for interpreting the quantitative classification results and illuminates the distinct visual characteristics of each preprocessing condition. Side-by-side comparisons of HQ, LQ, LQ+CLAHE, and LQ+DDPM images are presented for one representative case from each of the four disease categories: cataract, diabetic retinopathy, glaucoma, and normal.

In the LQ images, the effects of the medium-level degradation pipeline are clearly visible across all disease categories. Spatial resolution is markedly reduced relative to the HQ counterparts, resulting in softened vessel boundaries, reduced optic disc margin definition, and a general loss of fine structural detail. JPEG compression artifacts introduce subtle blocking patterns in homogeneous regions, and the combined effect of downsampling and Gaussian blur renders microaneurysms and intraretinal hemorrhages in the DR cases substantially less distinct. These visual observations are consistent with the quantitative performance gap between HQ and LQ conditions documented in Section 4.1.

After CLAHE preprocessing, local contrast is visibly enhanced across all cases, with vessel boundaries appearing sharper and the optic disc margin more clearly delineated than in the unprocessed LQ images. However, CLAHE's contrast stretching operates uniformly across image regions defined by its tile grid, occasionally introducing overamplification of noise and background texture in darker peripheral areas of the fundus. [25] This effect is particularly noticeable in images with uneven illumination, where CLAHE can produce an artificially mottled appearance that may introduce spurious texture features inconsistent with the training distribution. This observation provides a plausible visual explanation for the cross-fold variance observed in CLAHE's classification performance (F1 std = 0.0174): cases where illumination characteristics align favorably with CLAHE's parameters benefit from contrast enhancement, while others may be adversely affected by noise amplification, contributing to unstable performance across folds.

The DDPM-restored images present a qualitatively different appearance. Global illumination is more homogeneous than in either the LQ or LQ+CLAHE conditions, and the overall color tone and anatomical layout are well preserved. Vessel structures show improved visibility compared to the LQ baseline, and the optic disc region is rendered with greater clarity in several cases. Importantly, DDPM restoration does not introduce the localized contrast overamplification characteristic of CLAHE, resulting in a more uniform and naturalistic image appearance consistent with the learned fundus image prior. These visual properties align with the quantitative gains documented in Section 4.2, where DDPM-based restoration achieved meaningfully higher F1 and AUC than both the LQ baseline and CLAHE, alongside substantially lower cross-fold variance (F1 std = 0.0056). That said, fine pathological features such as microaneurysms remain difficult to discern in the restored images, reflecting the resolution constraints of the 128 × 128 DDPM training setup — a limitation that future work with higher-resolution conditional architectures may address.

DISCUSSION

5.1 Clinical implications

The present findings carry several clinical implications for automated retinal disease screening in resource-constrained settings. Our experiments demonstrate that moderate fundus image degradation leads to a consistent reduction of approximately 6 percentage points in F1-score and 1.8 points in AUC relative to high-quality imaging conditions — a gap that is both statistically robust and clinically meaningful. Critically, however, diffusion-based restoration recovers a substantial portion of this gap, achieving an F1 of 0.8807 compared to 0.8587 on

unenhanced LQ images, suggesting that software-level preprocessing can meaningfully compensate for hardware limitations in resource-constrained screening deployments. This finding is encouraging for teleophthalmology workflows in which fundus images are frequently captured with non-mydriatic cameras, variable illumination, or by operators without specialized training. [9]

At the same time, the observed performance gap between HQ and LQ conditions underscores the value of incorporating dedicated image enhancement modules into screening pipelines, particularly where hardware upgrades are not feasible. Our comparison between CLAHE and DDPM-based restoration offers a practically relevant finding: generative preprocessing not only achieves superior classification performance compared to conventional histogram-based enhancement, but does so with substantially improved stability across cross-validation splits. For deployment in heterogeneous clinical environments — where patient populations, camera models, and operator skill levels vary considerably — this reduction in performance variance may be as important as improvements in mean accuracy, as it suggests more predictable behavior when the system is exposed to imaging conditions not well-represented in the training data.

A further implication concerns the potential of diffusion-based restoration to standardize the visual appearance of fundus images acquired across diverse devices and protocols. By learning a generative prior over high-quality retinal structure, the DDPM acts as a structure-preserving denoiser that reduces background inhomogeneity and suppresses blur artifacts, potentially reducing the sensitivity of downstream classifiers to scanner-specific characteristics and illumination variability. The markedly reduced fold-to-fold variance (F1 std = 0.0056 for DDPM versus 0.0174 for CLAHE) suggests that diffusion-based preprocessing may contribute to more predictable and consistent performance when AI screening tools are deployed across new populations and imaging environments — a property of considerable practical value in large-scale public health programs.

5.2 Safety considerations (False Positive)

Any automated retinal classification system intended for clinical deployment must be evaluated not only in terms of aggregate accuracy but also with respect to its error profile, particularly false positives and false negatives, which carry direct implications for patient safety and resource utilization. In screening contexts, excessive false positive rates overload specialist referral pathways and reduce clinician trust, while elevated false negative rates risk delaying diagnosis of sight-threatening disease. [26] The consistently high AUC values across all four experimental conditions suggest a broadly acceptable balance between sensitivity and specificity. However, the fold-to-fold variability observed with CLAHE — an F1 standard deviation of 0.0174 compared with 0.0056 for DDPM — raises important safety considerations for real-world deployment.

A method whose performance fluctuates substantially depending on the training subset may produce unstable false positive rates when adapted to new clinical sites. The markedly lower variance with DDPM-based restoration implies a more consistent decision boundary across data partitions, which may translate into more stable and predictable false positive behavior. Qualitative inspection provides a plausible mechanistic explanation: CLAHE's localized contrast amplification can overemphasize vessel edges and background noise in poorly illuminated regions, potentially causing the classifier to overinterpret benign texture as pathological signal. In contrast, DDPM restoration draws processed images toward the learned fundus prior, suppressing noise more uniformly without introducing localized contrast artifacts.

It should be noted that this study did not include a systematic stratified analysis of false positive and false negative rates by disease category, which represents an important direction for future work. Prospective evaluation of referral-level safety metrics and calibration analysis will be necessary to fully characterize the safety profile of diffusion-based preprocessing in clinical deployment.

5.3 Limitations

This study has several limitations. First, experiments were conducted on a single publicly available fundus dataset with a fixed synthetic degradation protocol, which may not fully capture the diversity of real-world acquisition artifacts such as motion blur, lens contamination, or extreme underexposure. Second, the DDPM was trained under a resource-constrained configuration — 128 × 128 resolution and unconditional generation — rather than the conditional, high-resolution frameworks characteristic of state-of-the-art systems; conditional architectures with explicit LQ-to-HQ supervision may yield further classification gains beyond those observed here. Third, class-wise performance and calibration were not systematically analyzed, leaving disease-specific error patterns — particularly for glaucoma, which depends on subtle optic disc morphology, and diabetic retinopathy, which relies on microaneurysm detection sensitive to resolution loss — unexplored. Finally, all evaluation was conducted retrospectively on static datasets; prospective clinician-in-the-loop studies will be required to confirm whether the performance and stability advantages of diffusion-based preprocessing translate into reliable real-world screening outcomes.

6. CONCLUSION

This study examined whether diffusion-based image restoration can improve the robustness of automated retinal disease classification under low-quality imaging conditions representative of resource-constrained screening environments. Using a four-class fundus image dataset, we first established that synthetic medium-level degradation — simulating the acquisition artifacts of low-cost portable cameras — reduces classification F1-score by approximately 6 percentage points and AUC by 1.8 points relative to high-quality baselines, confirming that image quality exerts a measurable and consistent effect on AI-assisted ophthalmic screening performance.

Comparing CLAHE and DDPM-based restoration as preprocessing strategies, we found that their behavioral profiles differed substantially. DDPM-based restoration achieved meaningfully superior classification performance — recovering the majority of the LQ-induced performance gap and outperforming CLAHE by approximately 2 percentage points in F1 — while also exhibiting substantially lower cross-fold variance (F1 std = 0.0056 versus 0.0174 for CLAHE), pointing to improved generalization stability under identical resource constraints. Qualitatively, DDPM restoration produced more homogeneous illumination and clearer vessel and optic disc structures, whereas CLAHE occasionally introduced localized noise overamplification that may contribute to unstable classifier behavior across diverse imaging conditions.

These findings carry three main implications. First, classical histogram-based enhancement is not the optimal choice for robust retinal classification from degraded images, and generative restoration — even under significant resource constraints — can deliver both superior average performance and more stable generalization. Second, the advantages of DDPM in this setting span both accuracy and robustness, suggesting a compelling role for diffusion preprocessing in low-resource ophthalmic screening deployments where consistent performance across diverse imaging environments is essential. Third, this study provides an empirical characterization of conditions under which resource-constrained unconditional DDPM preprocessing delivers meaningful gains, offering a practical reference point for researchers and practitioners. Future work should investigate conditional diffusion frameworks, class-wise safety analysis, and prospective clinician-in-the-loop evaluation to further validate the clinical utility of generative preprocessing for equitable ophthalmic AI screening.

7. ACKNOWLEDGEMENT

This work was supported by the Korea Institute of Science and Technology Information (KISTI) and utilized the KISTI supercomputing infrastructure.

8. REFERENCES

[1] World Health Organization, *World Report on Vision*. Geneva, Switzerland: WHO, 2019.

[2] T. Y. Wong and C. M. G. Cheung, "Clinical overview of diabetic retinopathy and glaucoma," *Lancet Diabetes Endocrinol.*, vol. 10, no. 4, pp. 256–270, 2022.

[3] M. D. Abràmoff, M. K. Garvin, and M. Sonka, "Retinal imaging and image analysis," *IEEE Rev. Biomed. Eng.*, vol. 3, pp. 169–208, 2010.

[4] S. R. Bastawrous and A. R. Giardini, "Mobile eye screening in low-income settings: Opportunities and challenges," *Ophthalmology*, vol. 129, no. 5, pp. 521–530, 2022.

[5] T. Y. Wong et al., "Diabetic retinopathy," *Nat. Rev. Dis. Primers*, vol. 2, no. 16012, pp. 1–20, 2016.

[6] H. A. Quigley, "Glaucoma," *Lancet*, vol. 377, no. 9774, pp. 1367–1377, 2011.

[7] A. Sheeladevi, G. Lawrenson, C. F. Fielder, and A. Suttle, "Global prevalence of cataract and blindness: A systematic review and meta-analysis," *Eye*, vol. 33, no. 8, pp. 1210–1219, 2019.

[8] J. A. Lim et al., "Deep learning for automated detection of eye diseases from retinal fundus photographs," *npj Digit. Med.*, vol. 6, no. 1, p. 24, 2023.

[9] S. S. R. Bastawrous and A. R. Giardini, "Clinical applications and future directions of smartphone fundus imaging," *Clin. Exp. Ophthalmol.*, vol. 52, no. 2, pp. 123–134, 2024.

[10] M. Ahmed et al., "Smartphone-based fundus imaging–Where are we now?," *Ophthalmol. Retina*, vol. 7, no. 8, pp. 745–757, 2023.

[11] X. Li et al., "Assessing the impact of image quality on deep learning-based ophthalmic diagnosis," *Comput. Biol. Med.*, vol. 164, p. 107267, 2023.

[12] Y. Zhao et al., "DeepQuality improves infant retinopathy screening by joint image-quality assessment and enhancement," *Nat. Commun.*, vol. 14, no. 1, p. 6123, 2023.

[13] M. Rakib et al., "Eye disease classification using EfficientNet-B0 and data augmentation," *J. Ambient Intell. Humaniz. Comput.*, vol. 16, no. 5, pp. 2345–2357, 2025.

[14] A. Kazerouni et al., "Diffusion models for medical image restoration and enhancement: A systematic review," *Patterns*, vol. 5, no. 3, p. 100876, 2024.

[15] V. Gunavenkat Doddi, "Eye diseases classification," Kaggle dataset, 2022. [Online]. Available: https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification

[16] H. Li et al., "Deep learning-based classification of multiple fundus diseases using ultra-widefield images," *Front. Med.*, vol. 12, p. 1543210, 2025.

[17] S. Wang et al., “Multi-class eye disease classification using deep learning and EfficientNet-based fusion,” *Sci. Rep.*, vol. 16, no. 1, p. 1842, 2026.

[18] A. K. Patel, R. Huang, and P. Singh, “Enhancing fundus image-based diagnosis using CLAHE and deep CNNs,” *Biomed. Signal Process. Control*, vol. 93, p. 106040, 2024.

[19] Y. Zhang et al., “Comparative analysis of novel preprocessing techniques for fundus-based diabetic retinopathy grading,” *Med. Image Anal.*, vol. 92, p. 103046, 2025.

[20] J. Chung et al., “Diffusion models for medical image reconstruction,” *Br. J. Radiol. AI*, vol. 1, no. 1, p. ubae013, 2024.

[21] Y. Kim et al., “Diffusion model for medical image denoising, reconstruction and translation: A review,” *Comput. Biol. Med.*, vol. 184, p. 108102, 2025.

[22] S. Go et al., “Fundus image enhancement through direct diffusion bridges (FD3),” *Sci. Rep.*, vol. 14, no. 1, p. 16234, 2024.

[23] A. Kazerouni et al., “Diffusion models for image restoration and enhancement,” *arXiv preprint* arXiv:2308.09388, 2023.

[24] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Advances in Neural Information Processing Systems 33*, 2020.

[25] S. Singh et al., “CLAHE-CapsNet: Efficient retina OCT classification with contrast-limited adaptive histogram equalization,” *Transl. Vis. Sci. Technol.*, vol. 12, no. 11, p. 34, 2023.

[26] M. D. Abràmoff et al., “Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices,” *npj Digit. Med.*, vol. 1, no. 1, p. 39, 2018.