International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Multi-Class Cataract Detection from Pupil Images for Rural Communities

Anoushka Agrawal[1], Jaden Shiju[1], Saiprathist Reganti[1], Vivaan Rao[1]

Abstract

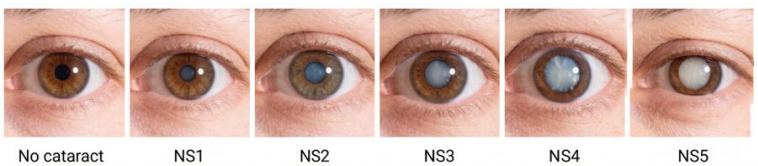
Cataracts are one of the leading causes of vision impairment worldwide, and their burden is disproportionately concentrated in low-resource settings such as rural communities. Many existing state-of-the-art deep learning models for cataract detection struggle to localize the clinically relevant region, the pupil and instead learn spurious correlations with irrelevant features such as skin color and eyebrow texture. In this work, it is proposed that cataract grading can be improved by restricting model input to the pupil region and addressing two systematic data quality issues: scale variability and specular reflection artifacts. We developed a two-path preprocessing pipeline that applies partial convolution-based inpainting to remove bright spot artifacts and randomized zoom augmentation via a MONAI pipeline to normalize scale. A ResNet-18 classifier is then trained on pupil-cropped images to perform six-class cataract grading (No Cataract, NS1–NS5). Experiments on 1,639 pupil images demonstrate that the combined training strategy, mixing inpainted and non-inpainted images with data augmentation, achieves the best multi-class AUC of 0.8919. It is observed that the model reliably distinguishes healthy from cataractous eyes, while confusion between adjacent intermediate grades (NS2–NS3) remains a challenge for future work.

1. Introduction

A cataract is the clouding of the crystalline lens of the eye caused by the aggregation and degradation of lens proteins and fibers. This clouding scatters incoming light before it reaches the retina, resulting in blurry, dim, or faded vision. Cataracts may arise from aging, ocular trauma, genetic predisposition, environmental factors, or systemic conditions such as diabetes. They typically develop in both eyes of an individual, although the rates of progression may be independent.

Severity is clinically graded using systems such as the Nuclear Sclerosis (NS) scale, which ranges from NS1 (mild clouding) to NS5 (severe, dense clouding). Figure 1 depicts the appearance of the different levels of cataract. When the cataract is detected at an early stage and has minimal impact on functional vision, management may be limited to corrective eyeglasses. In cases where vision deterioration significantly impairs daily activity, surgical intervention is required.

Figure 1: Levels of Cataract



Despite the availability of effective treatment, cataract remains the leading cause of preventable blindness globally, particularly in low- and middle-income countries where specialist ophthalmological care is scarce. Automated cataract detection from smartphone or fundus images could substantially expand access to diagnosis in rural settings. However, existing deep learning approaches suffer from a fundamental interpretability flaw: rather than attending to the lens where the cataract physically manifests, models trained on full eye images have been shown to exploit confounding features such as skin tone, periorbital texture, and eyebrow morphology, as revealed by Class Activation Map (CAM) visualizations. This limits the clinical validity and generalizability of such models.

Figure 2: Grad-CAM Visualization

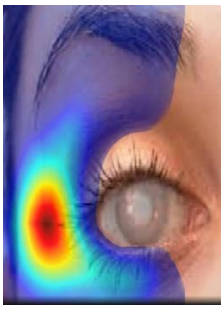
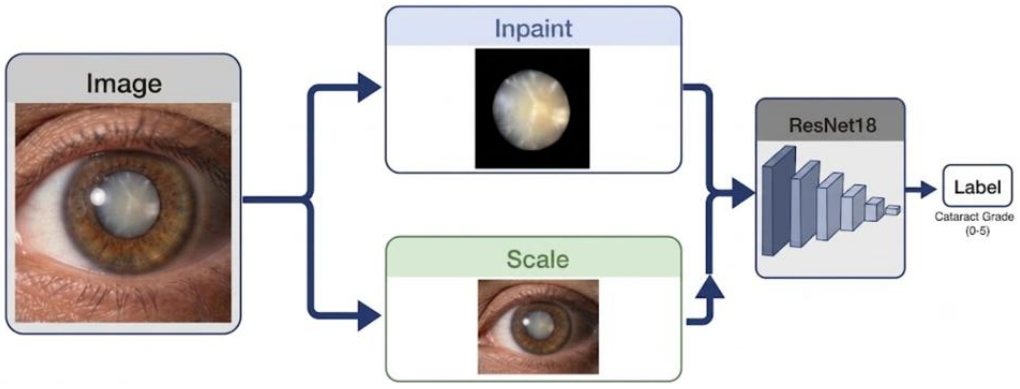


Figure 2 demonstrates a Grad-CAM Visualization focusing on skin tone rather than the pupil. In this work, a system named Mo-Cat is presented, which addresses this shortcoming through a three-stage pipeline: (1) pupil detection and extraction to constrain the model's field of view to the clinically relevant region; (2) artifact removal via partial convolution inpainting and scale normalization via zoom augmentation; and (3) six-class cataract grade classification using a ResNet-18 backbone. Figure 3 displays the final two stages of the pipeline we implemented, in which the model uses inpainting technology and zoom augmentation paired with a ResNet-18 backbone to output a label.

Figure 3: Outcome Flow Chart



We propose a pupil-centric preprocessing framework that removes brightness artifacts through partial convolution inpainting, enabling cleaner and more consistent feature extraction from clinical pupil images. To further improve model generalization, we employ a multi-level data augmentation strategy using MONAI that addresses scale variability commonly encountered in clinical pupil images. Through extensive experimentation, we demonstrate that training on a combination of both inpainted and non-

inpainted images yields the most robust multi-class classifier, achieving an AUC of 0.8919 and outperforming models trained exclusively on either data type.

2. Related Work

Ki Young Son et al. developed and validated a deep learning-based cataract detection and grading system using Slit-Lamp and Retro-Illumination Photographs, grading severity using the Lens Opacities Classification System III (LOCS III), and published their findings in Ophthalmology Science. Their work demonstrated the feasibility of automated lens opacity grading from clinical imagery but relied on controlled slit-lamp acquisition settings not available in low-resource environments. [3]

Amr Elsawy et al. proposed DeepOpacityNet, a deep learning network for cataract detection from color fundus photography (CFP) images, published in Communications Medicine. Their approach divided the data by cataract type and demonstrated competitive detection accuracy on publicly available fundus datasets. However, fundus photography requires dedicated equipment and trained operators, limiting its applicability for community-level screening. [4]

In contrast to these prior works, the present study targets smartphone-captured pupil images, explicitly addresses the scale and reflectance artifacts endemic to such data, and demonstrates that confining the model's receptive field to the pupil region is critical for clinically meaningful feature learning.

3. Problem Formulation

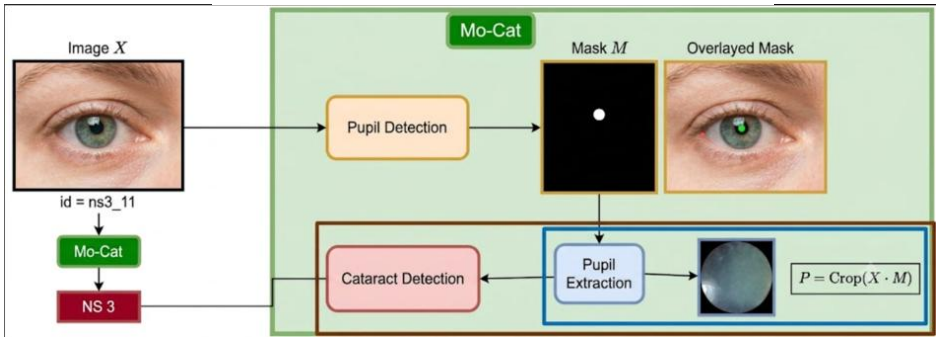
Given a raw eye image X, the objective is to predict one of six cataract severity labels y ∈ {No Cataract, NS1, NS2, NS3, NS4, NS5}. The pipeline first extracts the pupil region P = Crop (X · M), where M is a binary mask localizing the pupil. The extracted pupil image P (or its inpainted counterpart Ô) is then passed to a classifier fθ that returns the grade label. Two systematic image quality challenges complicate direct classification: Scale variability: Pupil images are captured at varying distances, resulting in pupils of different relative sizes. Without normalization, models may conflate scale-induced differences in feature sharpness with genuine severity differences, leading to grade underestimation or grade merging. Specular reflection artifacts: Camera flash and ambient lighting produce bright spot corruption on the lens surface. These high-intensity regions are not diagnostic features; however, they dominate local pixel statistics and can mislead the classifier into attending to brightness rather than cloudiness.

4. Methodology

4.1 Pupil Detection and Extraction

Pupil detection is performed using a combination of HSV-space color masking and Hough circle detection implemented with OpenCV, followed by refinement using a MONAI UNet segmentation model trained on the dataset's pupil annotations. The binary mask M is applied to the original image X via element-wise multiplication, as illustrated in Figure 4, and the bounding box of the non-zero mask region is used to crop the pupil patch P. All cropped patches are resized to a fixed spatial resolution prior to downstream processing.

Figure 4: Project Layout

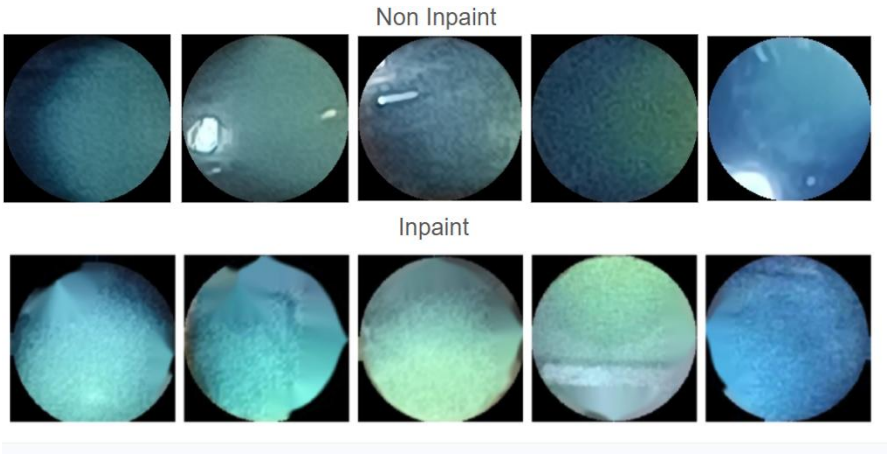


4.2 Inpainting for Artifact Removal

Specular reflections and glare are detected as high-intensity regions within the pupil mask. Artifact-corrupted pixels are identified by thresholding on pixel intensity. A binary corruption mask (where 1 indicates invalid/corrupted pixels) is constructed and passed along with the corrupted pupil image X^m to a partial convolution inpainting method [1] to obtain the reconstructed output $\hat{O} = f_\theta(X^m, M)$

This process effectively removes artifacts such as bright spots, reflections, and glares while preserving lens morphology. Partial convolution operates in three sequential steps. First, masked convolution is applied: the convolution kernel operates only on valid pixels (M = 0), ignoring corrupted regions. Second, renormalization is applied: the output is scaled by the inverse proportion of valid pixels within the kernel's receptive field, ensuring that the effective contribution of valid pixels is preserved regardless of mask density. Third, mask update is performed: at each layer, the corruption mask is eroded so that pixels predicted by prior layers are progressively incorporated as valid in subsequent layers. This progressive mask shrinkage enables the network to fill in corrupted regions with contextually consistent reconstructions that preserve the pupil's surface morphology. In Figure 5, the result of the partial convolution inpainting method on the pupils used in our research is shown. Bright spots and other similar aberrations are removed from the pupil, ensuring more accurate categorization.

Figure 5: Inpainting Results



4.3 Scale Normalization via Data Augmentation

To address scale variability, a MONAI [2] randomized zoom augmentation pipeline is applied during training. Random zoom factors are sampled uniformly from a predefined range, simulating the full spectrum of capture distances observed in the dataset. Crucially, the output spatial resolution is held constant after zooming, so the classifier always receives fixed-size inputs regardless of the simulated capture distance. Additional augmentations applied include random rotation, Gaussian blur, and brightness normalization, all randomized at each training iteration to improve generalization. These transformations ensure the model is robust to variations in capture distance.

4.4 Classification Network

A ResNet-18 backbone, pretrained on ImageNet, is used as the feature extractor. The final fully connected layer is replaced with a six-class linear classifier corresponding to the labels {No Cataract, NS1, NS2, NS3, NS4, NS5}. The model is trained using categorical cross-entropy loss with the Adam optimizer. Early stopping with a patience of 50 epochs is applied to prevent overfitting, with a maximum training budget of 200 epochs.

4.5 Training Configurations

To isolate the individual and combined contributions of inpainting and data augmentation, we evaluate four distinct training configurations. The first configuration, Non-Inpaint without Data Augmentation (DA), serves as a baseline classifier trained on raw pupil crops. The second, Non-Inpaint with DA, applies augmentation techniques to the raw pupil crops during training. The third configuration, Inpaint with DA, trains exclusively on inpainted pupil crops while incorporating data augmentation. Finally, the All Together (AT) configuration combines both inpainted and non-inpainted images within the training set and is evaluated both with and without data augmentation to assess the benefits of data diversity and preprocessing jointly.

5. Experimental Setup

A dataset of 1,639 unique pupil images adapted from the dataset is used, divided into a 60/20/20 split for training, validation, and testing, respectively. The six classes are: No Cataract, NS1, NS2, NS3, NS4, and NS5, spanning mild to severe lens opacity. The primary evaluation metric is the multi-class Area Under the Receiver Operating Characteristic Curve (AUC), which measures how accurately the model ranks different cataract severity levels. An AUC of 1.0 indicates perfect ranking performance. All experiments are conducted on the same data split to ensure fair comparison.

6. Results

Table I summarizes the AUC scores achieved by all training configurations.

Table 1: AUC Results Across Training Configurations

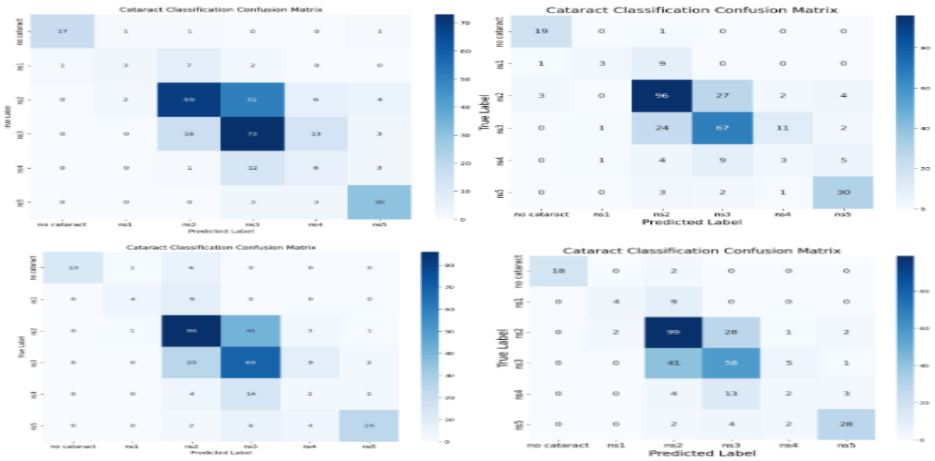
Model Configuration	AUC

Non-Inpaint, without DA	0.8549
Non-Inpaint, with DA	0.8716
Inpaint, without DA	0.8583
Inpaint, with DA	0.8738
All Together, without DA	0.8000
All Together, with DA (Best)	0.8919

It is observed that data augmentation consistently improves performance within each inpainting condition. The best-performing configuration is All Together with Data Augmentation (AUC = 0.8919). Training on a mixture of inpainted and non-inpainted data, combined with augmentation, makes the model simultaneously robust to brightness artifacts and scale variability, yielding the highest generalization. Conversely, All Together without augmentation achieves the lowest AUC (0.8000), indicating that the beneficial effect of mixed training depends on augmentation diversity to prevent the model from overfitting to the training distribution.

Figure 6 models for four confusion matrix variants: **(a)** Inpaint with Data Augmentation (top-left), **(b)** Non-Inpaint with Data Augmentation (top-right), **(c)** Combined with Data Augmentation (bottom-left), and **(d)** Combined without Data Augmentation (bottom-right).

Figure 6: Confusion Matrix



Confusion matrix analysis reveals that the model reliably distinguishes healthy eyes (No Cataract) from eyes with any degree of cataract, with minimal misclassification across this boundary. The primary remaining difficulty lies in separating NS2 from NS3, which represents adjacent and visually similar intermediate grades.

Figure 7: Grad-CAM Visualizations of Pupil Regions Under Four Training Settings: (A) Non-Inpaint With Data Augmentation (Top), (B) Inpaint with Data Augmentation (Second), (C) All Together With Data Augmentation (Third), And (D) All Together Without Data Augmentation (Bottom)

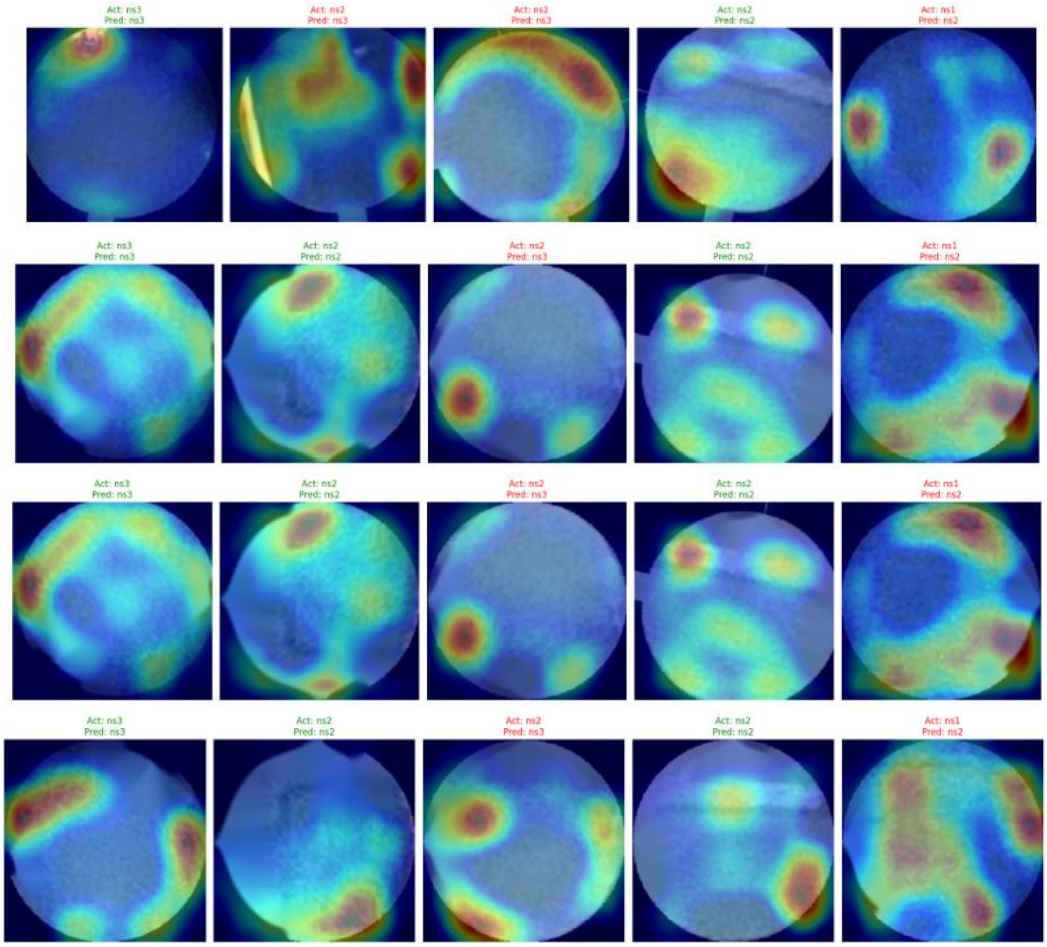


Figure 7 shows that, while the model generally attends to the pupil region, attention patterns are not yet uniform across the pupil surface, suggesting that further interpretability-driven regularization could improve both accuracy and clinical trustworthiness.

7. Discussion

The results confirm that the proposed pupil-centric, artifact-aware training strategy is effective for multi-class cataract grading. The superiority of the All Together configuration supports the hypothesis that models trained jointly on clean and corrupted data learn more generalizable representations than models trained on either type exclusively. This mirrors established practices in robust machine learning, where training distribution diversity is a key driver of test-time robustness.

The NS2–NS3 confusion is expected given the subtle visual difference between these grades, even for trained ophthalmologists. Addressing this boundary may require either higher-resolution input, additional clinically derived features (e.g., spectral reflectance profiles), or ordinal regression losses that explicitly encode severity ordering.

A notable limitation is the non-uniform attention patterns observed in the CAM visualizations. Ideally, attention should be distributed across the entire pupil surface, since cataract clouding is a global lens

property. Incorporating attention supervision or spatial regularization losses informed by clinical knowledge of cataract morphology represents a promising direction for future work.

8. Limitations and Future Work

The inpainting module does not yet eliminate all specular artifacts; some edges and corners of the reconstructed pupil region exhibit visible boundary artifacts that could still mislead the classifier. CAM-based analysis indicates that attention is not uniformly distributed over the pupil. Future work will investigate attention supervision to enforce clinically meaningful spatial focus. Differentiation between NS2 and NS3 remains the primary bottleneck. Ordinal classification losses, higher-resolution imaging, or additional imaging modalities (e.g., slit-lamp imagery) may improve intergrade discrimination. Expansion of the dataset and further augmentation strategies are planned to improve robustness and inter-class balance.

9. Conclusion

A pupil-centric multi-class cataract grading system is presented that addresses two key data quality challenges: specular reflection artifacts and scale variability through partial convolution inpainting and MONAI-based zoom augmentation. It is demonstrated that training a ResNet-18 classifier on a mixture of inpainted and non-inpainted pupil images with data augmentation achieves the best multi-class AUC of 0.8919, substantially outperforming single-condition baselines. The approach directly addresses the shortcoming of existing models that attend to clinically irrelevant image regions, offering a more interpretable and deployable cataract screening tool for low-resource clinical settings.

10.Acknowledgements

The authors gratefully acknowledge LEF Hospital for providing valuable insights into the healthcare challenges faced by rural communities. We also thank the creators and maintainers of the available image dataset used to train and validate the deep learning model presented in this study. Finally, we express our gratitude to all collaborators involved in data handling, annotation, and analysis.

References

1. Liu, Guilin, et al. "Image inpainting for irregular holes using partial convolutions." Proceedings of the European conference on computer vision (ECCV). 2018.

2. Cardoso, M. Jorge, et al. "Monai: An open-source framework for deep learning in healthcare." arXiv preprint arXiv:2211.02701 (2022).

3. Son, Ki Young, et al. "Deep learning-based cataract detection and grading from slit-lamp and retro-illumination photographs: model development and validation study." Ophthalmology Science 2.2 (2022): 100147.

4. Elsawy, Amr, et al. "A deep network DeepOpacityNet for detection of cataracts from color fundus photographs." Communications Medicine 3.1 (2023): 184.