E-ISSN: 3025-0501
Vol. 3, No. 6, November 2025
pp. 315 - 331
DOI: https://doi.org/10.57185/qkc5mc07

Almakki Health Informatics Journal

Image Classification of Eye Diseases Using MobileNetV3-Large with CLAHE Contrast Enhancement

Zahra Zhafirah
Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) IKMI Cirebon, Indonesia
Corresponding Author: zahrazfrh732@gmail.com

Keywords:
CLAHE; MobileNetV3-Large; image classification; eye diseases; transfer learning; confusion matrix.

This is an open access article under the CC BY-SA license.

Copyright © 2026 by Author. Published by Al-Makki Publisher

ABSTRACT

Background: Deep learning–based image classification of eye diseases has the potential to support initial screening by recognizing visual patterns in images. However, image illumination and contrast can affect the features learned by the model.
Objective: This study aimed to characterize the performance of MobileNetV3-Large on eye image data processed using Contrast Limited Adaptive Histogram Equalization (CLAHE) and analyze class-specific classification errors.
Methods: Images from the Eye Diseases Classification dataset (cataract, diabetic retinopathy, glaucoma, and normal) were resized to 224 × 224 pixels, and CLAHE was applied to the luminance channel of the LAB color space. Data were stratified into training, validation, and test sets (64:16:20). The ImageNet-pretrained model was trained for 30 epochs, and the final-epoch model was evaluated.
Results: Of 844 test images, 768 were correctly classified (accuracy 91.00%; 95% confidence interval: 88.87–92.75%; macro-F1 90.79%). Recall was highest for diabetic retinopathy (99.55%) and lowest for normal images (76.28%), which were most often misclassified as glaucoma. The training curves suggest possible overfitting.
Conclusion: Because no baseline without CLAHE was evaluated, these results characterize the MobileNetV3-Large + CLAHE pipeline but do not establish that CLAHE improved classification performance. A non-CLAHE baseline, duplicate-image screening, validation-based checkpoint selection, and external validation are required before any consideration of clinical use.

INTRODUCTION

Eye diseases can affect a person's quality of life and ability to perform daily activities. Fundus images and other ophthalmic images provide visual information that can support the identification of specific conditions. In recent years, deep learning has become an important approach to medical image analysis because it can learn visual representations directly from data. A review by Litjens et al. (2017) describes the broad development of deep learning for various medical image analysis tasks, while Shen et al. (2017) discuss the potential and challenges of these approaches in biomedical applications.

At least 2.2 billion people worldwide are estimated to live with vision impairment or blindness, of whom at least 1 billion have a condition that could have been prevented or has yet to be addressed; cataract, glaucoma, and diabetic retinopathy are among the principal causes

(Li et al., 2024; Sah et al., 2026). In Indonesia, the Rapid Assessment of Avoidable Blindness, which estimated the prevalence and causes of blindness among people aged 50 years and older, identified cataract as the leading cause of blindness (Rahayu et al., 2023). Because many of these conditions are treatable when detected early, image-based tools that assist the triage of eye images could help prioritize referrals, particularly where access to eye-care specialists is limited. Such tools, however, require rigorous validation, because strong internal test performance does not automatically translate into clinical impact (Kelly et al., 2019).

The performance of a classification model is not determined solely by its network architecture. Variations in illumination, contrast, sharpness, image acquisition quality, and class distribution can affect the visual patterns available to the model. In fundus imaging, local details, such as blood vessels and small lesions, can be difficult to discern when an image is poorly illuminated or has low contrast. Therefore, image enhancement before training is a widely investigated preprocessing approach. Alwazzan et al. (2021) evaluated a fundus image enhancement method combining a Wiener filter and CLAHE, while Sonali et al. (2019) investigated the use of denoising and CLAHE for contrast enhancement in fundus images.

Contrast Limited Adaptive Histogram Equalization (CLAHE) is a local contrast enhancement method that limits histogram amplification to prevent excessive contrast enhancement. Unlike global histogram equalization, CLAHE operates on small regions of an image and combines the results through interpolation. The application of CLAHE to medical imaging must be evaluated empirically because visual enhancement does not necessarily translate into improved model accuracy. Hayati et al. (2023), for example, reported improved classification accuracy with CLAHE for some architectures used to classify diabetic retinopathy but found that ResNet34 performed better without CLAHE. These findings indicate that the effects of preprocessing depend on the model architecture and dataset.

This study used MobileNetV3-Large, a convolutional neural network (CNN) architecture designed for computational efficiency through a combination of efficient convolutional blocks, squeeze-and-excitation modules, and optimized activation functions suitable for resource-constrained devices (Howard et al., 2019). The model was initialized with pretrained ImageNet weights, and its final classification layer was customized to produce predictions for four classes. Transfer learning was expected to help the model leverage previously learned visual representations when trained on the target dataset.

Deep learning, particularly convolutional neural networks (CNNs), has been used for the classification, detection, segmentation, and quantitative analysis of medical images. CNNs can learn hierarchical features, ranging from edges and textures to more complex visual patterns. However, performance on internal test data does not automatically guarantee generalizability to different populations, cameras, clinical facilities, or image acquisition protocols. Litjens et al. (2017) and Shen et al. (2017) emphasize that data evaluation, validation, and interpretation are important components of developing medical image analysis systems.

MobileNetV3 is a family of CNN architectures that prioritizes computational efficiency through inverted residual blocks, squeeze-and-excitation modules, and optimized activation functions. Howard et al. (2019) introduced MobileNetV3 as an architecture developed using neural architecture search and optimization techniques for mobile applications. In this experiment, pretrained ImageNet weights were used to initialize the model. The final

classification layer was replaced with a linear layer containing four output units, corresponding to the four classes in the dataset.

Transfer learning is useful when the target dataset is smaller than the pretraining dataset. However, domain differences between ImageNet images and medical images may limit the benefits of the pretrained representations. Because all model parameters were trained in this experiment, the final performance was influenced by the dataset size, data augmentation, learning rate, and training duration. Future research could compare full fine-tuning with partial freezing of the backbone, although such a comparison falls outside the scope of the experiments reported in this study.

Yosinski et al. (2014) showed that the transferability of features learned from large natural-image collections such as ImageNet (Deng et al., 2009) decreases as the dissimilarity between the source and target tasks increases, whereas fine-tuning the transferred layers can partly compensate for this gap. This finding motivates the full fine-tuning strategy adopted in the present experiment. CLAHE divides an image into local tiles and performs histogram equalization with a clipping limit to reduce excessive contrast enhancement. The parameters used in the notebook were clipLimit = 2.0 and tileGridSize = (8, 8). Preprocessing was performed after resizing each image to 224 × 224 pixels. The BGR image was converted to the LAB color space, CLAHE was applied to the L channel (lightness), and the L, a, and b channels were then recombined before the image was converted back to BGR.

The application of CLAHE to fundus images has been investigated in various contexts involving image quality enhancement and classification. Alwazzan et al. (2021) evaluated a hybrid approach combining Wiener filtering and CLAHE to enhance structures in retinal images. Hayati et al. (2023) compared several architectures for diabetic retinopathy classification and demonstrated that the effects of CLAHE varied across models. Therefore, claims about the benefits of CLAHE should be supported by comparative experiments using equivalent datasets, models, and training procedures.

CLAHE originated from adaptive histogram equalization and was subsequently formalized as a contrast-limited variant (Härtinger & Steger, 2024). In retinal imaging, CLAHE has been applied alone (Wanling & Shah, 2025) and in combination with other enhancement techniques, including a Retinex-based variational model with CLAHE in the CIELUV color space (Huang et al., 2020) and CLAHE with ESRGAN-based super-resolution for diabetic retinopathy prediction (Alwakid et al., 2023).

This experiment classified eye images into four categories: cataract, diabetic retinopathy, glaucoma, and normal. These categories represent the labels assigned in the dataset and should not be interpreted as clinically verified diagnoses established by this study. The images and labels were obtained from the publicly available Eye Diseases Classification dataset (Alsohemi & Dardouri, 2025; Sah et al., 2026), whose provenance is described in the Method section. Accuracy is the proportion of all predictions that are correct. Precision measures the proportion of predictions assigned to a particular class that are correct, whereas recall measures the proportion of actual samples belonging to that class that the model correctly identifies.

The F1-score is the harmonic mean of precision and recall. In multiclass classification, macro-averaging calculates the unweighted mean of a metric across all classes, regardless of their respective sample counts. This approach provides a more balanced view of class-level performance than accuracy alone, particularly when the dataset is imbalanced. The confusion

matrix displays actual classes in rows and predicted classes in columns, allowing misclassification patterns to be examined directly. Previous studies on the same four categories have reported varying results. Agustini et al., (2026) compared MobileNetV2, Xception, and EfficientNet-B0 on 4,217 fundus images and reported a mean five-fold cross-validation accuracy of 0.85 for EfficientNet-B0, together with a marked decline in accuracy on external clinical images.

Addyna et al., (2026) reported an accuracy of 89.17% for an optimized MobileNet on 1,200 Kaggle fundus images, whereas Alsohemi and Dardouri (2025) reported an accuracy of 95.12% for EfficientNetB3 on the same public dataset. None of these studies combined MobileNetV3-Large with CLAHE applied to the luminance channel, and Hayati et al. (2023) examined CLAHE only for diabetic retinopathy. The novelty of the present study therefore lies in a transparent and reproducibly documented evaluation of a MobileNetV3-Large + CLAHE pipeline for four-class eye disease classification, supported by confidence intervals and a class-level error analysis that explicitly delimits what can and cannot be inferred about the effect of CLAHE.

The study aimed to characterize the performance of MobileNetV3-Large on eye image data processed using CLAHE and analyze class-specific classification errors. It did not evaluate clinical diagnostic performance, segment lesions, or experimentally compare CLAHE-enhanced images with images processed without CLAHE. This limitation is important to ensure that the conclusions remain within the scope of the evidence generated by the experiment. The objectives of this study were to measure model performance on the test dataset, analyze precision, recall, F1-score, and the confusion matrix for each class, and identify limitations in the training process that should be addressed in future research.

The main contributions are an empirical evaluation of the CLAHE and MobileNetV3-Large configuration, an analysis of interclass classification errors, and the identification of additional validation steps needed in subsequent studies. These contributions are specific to the dataset and experimental configuration used and do not constitute evidence of clinical validation.

METHOD

This study employed a quantitative, computational experimental design. One primary configuration was evaluated, consisting of CLAHE preprocessing followed by MobileNetV3-Large training. The notebook did not include a baseline group without CLAHE; therefore, the study evaluated a single configuration rather than conducting a causal experiment to measure the performance improvement attributable to CLAHE.

Dataset source, image-level identifiers, patient-level grouping, annotation procedures, and licensing information were not available in the manuscript materials. Consequently, independence between subsets and the absence of duplicate or near-duplicate images could not be verified (Kapoor & Narayanan, 2023). These omissions limit reproducibility and may affect the interpretation of internal test performance; no unverified provenance or leakage-check results are asserted here.

The notebook loaded images from the eye_dataset directory and grouped the files according to their class subdirectories. Readable images with JPG, JPEG, PNG, or BMP extensions were included in the dataset inventory. The dataset was split using train_test_split,

with stratification based on class labels and random_state=42. The first split allocated 20% of the data to testing. Of the remaining 80%, 20% was allocated to validation. Consequently, the final dataset proportions were approximately 64% for training, 16% for validation, and 20% for testing.

According to the class distribution graph, the test set contained 844 images: 208 labeled as cataract, 220 as diabetic retinopathy, 201 as glaucoma, and 215 as normal. The graph implied a total dataset size of 4,217 images. However, the numbers of training and validation images shown in the graph needed to be verified against the output table from the original notebook before manuscript submission, because the graph provided only a visualization and its axis labels were rounded.

Each image was resized directly to 224 × 224 pixels using linear interpolation. This operation did not preserve the original aspect ratio and could therefore have introduced geometric distortion. This limitation was acknowledged in the study. After resizing, each image was converted from the BGR color space to the LAB color space. CLAHE, with a clipLimit of 2.0 and a tile grid size of 8 × 8, was applied to the luminance channel (L). The image was then converted back to BGR and stored as a preprocessed image.

During data loading, OpenCV was used to read each image, which was then converted from BGR to RGB. The training transformations consisted of ToTensor, RandomHorizontalFlip with a probability of 0.5, and normalization using the ImageNet mean and standard deviation. The validation and test sets underwent ToTensor and the same normalization procedure without random augmentation. The appropriateness of horizontal flipping needed to be considered to ensure that it did not alter clinically relevant image orientation or the interpretation of the target labels.

Training Architecture and Parameters

Reporting clarification: the training-parameter table identifies the optimizer as Adam (Reyad et al., 2023), whereas the abstract and conclusion identify AdamW (Zhou et al., 2024). The original executable notebook or training log must be checked to resolve this inconsistency before submission; the optimizer cannot be independently confirmed from the manuscript alone.

Table 1. Training architecture and parameter configuration

Components	Logged configurations
Models	MobileNetV3-Large pretrained ImageNet
Classifier	The final linear layer is replaced into 4 outputs
Input size	224 × 224 pixels
Loss	CrossEntropyLoss
Optimizer	Adam
Learning rate	0,0001
Weight decay	0,01
Batch size	32
Epoch	30
Seed	42
Increased training	RandomHorizontalFlip, p = 0,5
Normalization	Mean [0,485; 0,456; 0,406], SD [0,229; 0,224; 0,225]

Checkpoint/early stopping	Not visible on notebooks; final model after 30 epochs evaluated

The training DataLoader used drop_last=True, whereas the calculations of training loss and accuracy were normalized by the size of the entire training dataset. Because the final incomplete batch was excluded from training, this discrepancy could have affected the reported training metrics. For subsequent experiments, setting drop_last=False would allow all training samples to be included in each epoch, unless excluding incomplete batches was methodologically justified. The loss and accuracy calculations should also be aligned with the actual number of samples processed.

The evaluation used a single stratified split and random seed. Because no confidence intervals, repeated runs, or external dataset evaluations were reported, performance estimates are descriptive and should not be generalized beyond the assessed test subset. No inferential comparison between preprocessing conditions was possible. The experimental workflow consisted of the following steps: (1) loading the dataset; (2) resizing the images and applying CLAHE; (3) splitting the data using stratified sampling; (4) constructing the Dataset and DataLoader objects; (5) loading the ImageNet-pretrained MobileNetV3-Large model; (6) replacing the classifier to accommodate four classes; (7) training the model for 30 epochs; (8) recording training and validation loss and accuracy; and (9) generating predictions on the test set and presenting the classification report and confusion matrix.

The analysis was descriptive. The confusion matrix was used to recalculate accuracy, precision, recall, and F1-score for each class. Macro-averaged metrics were calculated as the arithmetic means of the corresponding metrics across the four classes. The training and validation loss and accuracy curves were examined to assess training behavior and potential generalization issues. No statistical significance tests or confidence interval analyses were performed. Therefore, observed differences in the evaluation metrics were not interpreted as statistically significant differences.

RESULTS AND DISCUSSION

Data Distribution

The distribution graph shows that all four classes are represented in the training, validation, and test subsets. The confusion matrix contains 844 test samples, distributed as follows: 208 cataract images, 220 diabetic retinopathy images, 201 glaucoma images, and 215 normal images. The class distribution in the test subset is relatively balanced, with no extreme imbalance among the four classes. The exact counts reported in Table 1 confirm that the stratified split preserved the class proportions of the original dataset (24.6% cataract, 26.0% diabetic retinopathy, 23.9% glaucoma, and 25.5% normal).

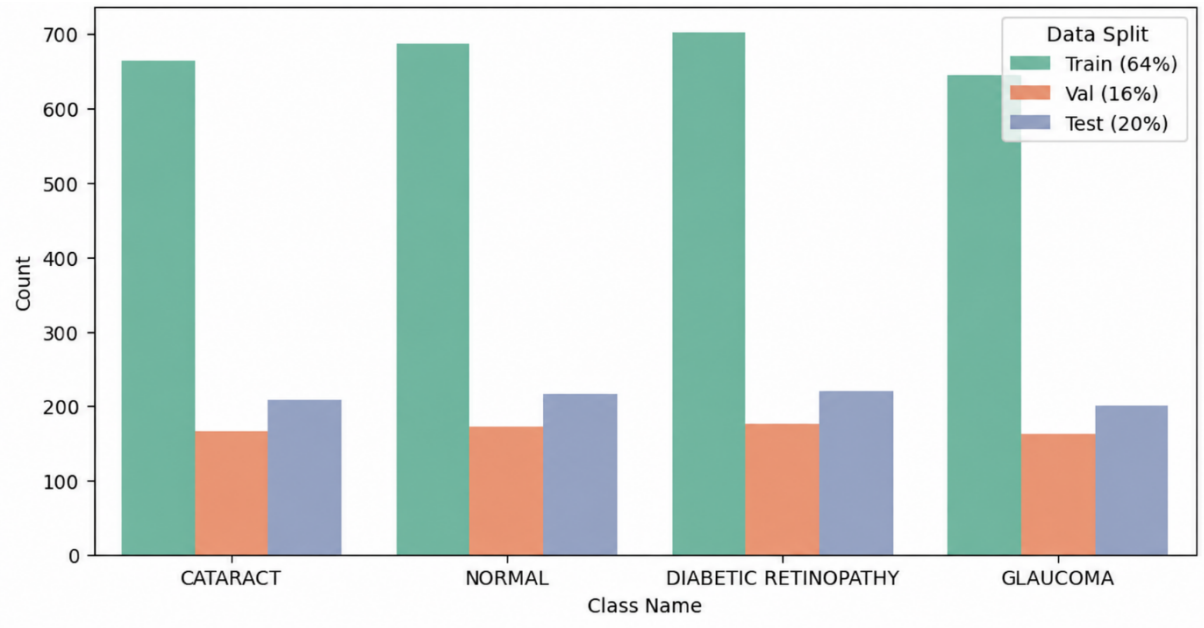


Figure 1. Distribution of images per class in training, validation, and testing data

Loss and Accuracy Curve

The training curves show a rapid decrease in training loss toward zero, while training accuracy increases to nearly 99%. In contrast, validation loss decreases during the initial epoch but subsequently fluctuates and increases sharply in later epochs. Validation accuracy remains approximately 89–93% after the initial phase and does not exhibit a corresponding increase comparable to that of training accuracy. This pattern suggests possible overfitting: the model continues to fit the training data increasingly well but does not demonstrate consistent improvement on the validation data.

Visual inspection of Figure 2 indicates that the validation loss reached its minimum early in training, at approximately the sixth epoch, and increased thereafter; the exact epoch is not reported because it cannot be confirmed from the figure alone. Accordingly, this study does not claim that any particular epoch represents the optimal checkpoint, and the reported test results correspond to the final model after 30 epochs rather than to a validation-selected model. Because selecting a model on the basis of test-set performance would introduce evaluation bias, checkpoint selection should rely exclusively on validation loss.

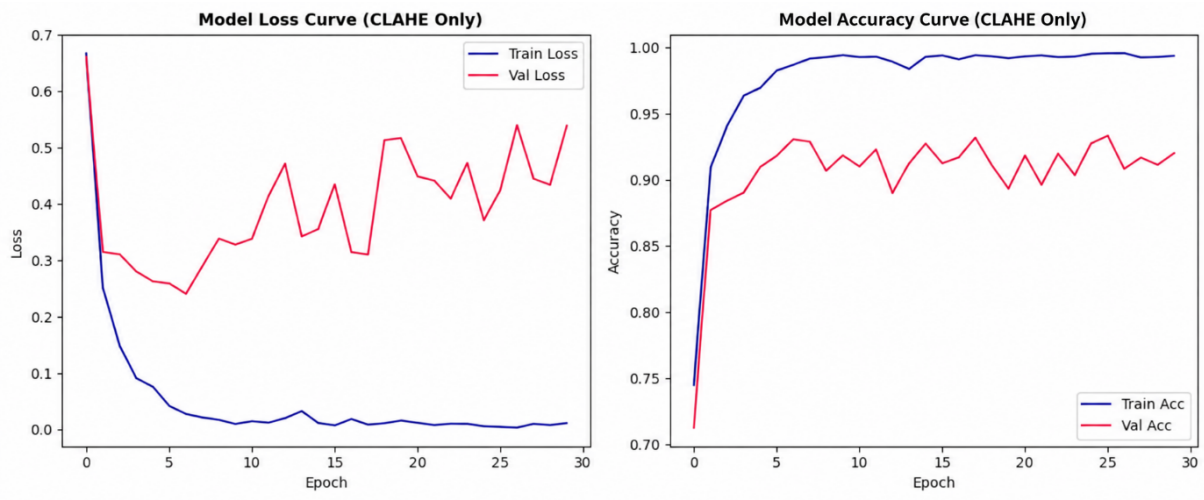


Figure 2. Loss curve and training-validation accuracy for 30 epochs

Confusion Matrix

The confusion matrix shows 768 correct predictions among 844 test images. The largest source of error was the misclassification of normal images as glaucoma, accounting for 34 images. In addition, 10 glaucoma images were misclassified as normal, nine cataract images were misclassified as glaucoma, and nine normal images were misclassified as diabetic retinopathy. This pattern suggests that the distinction between certain classes, particularly normal and glaucoma, is a major source of classification error in the test set.

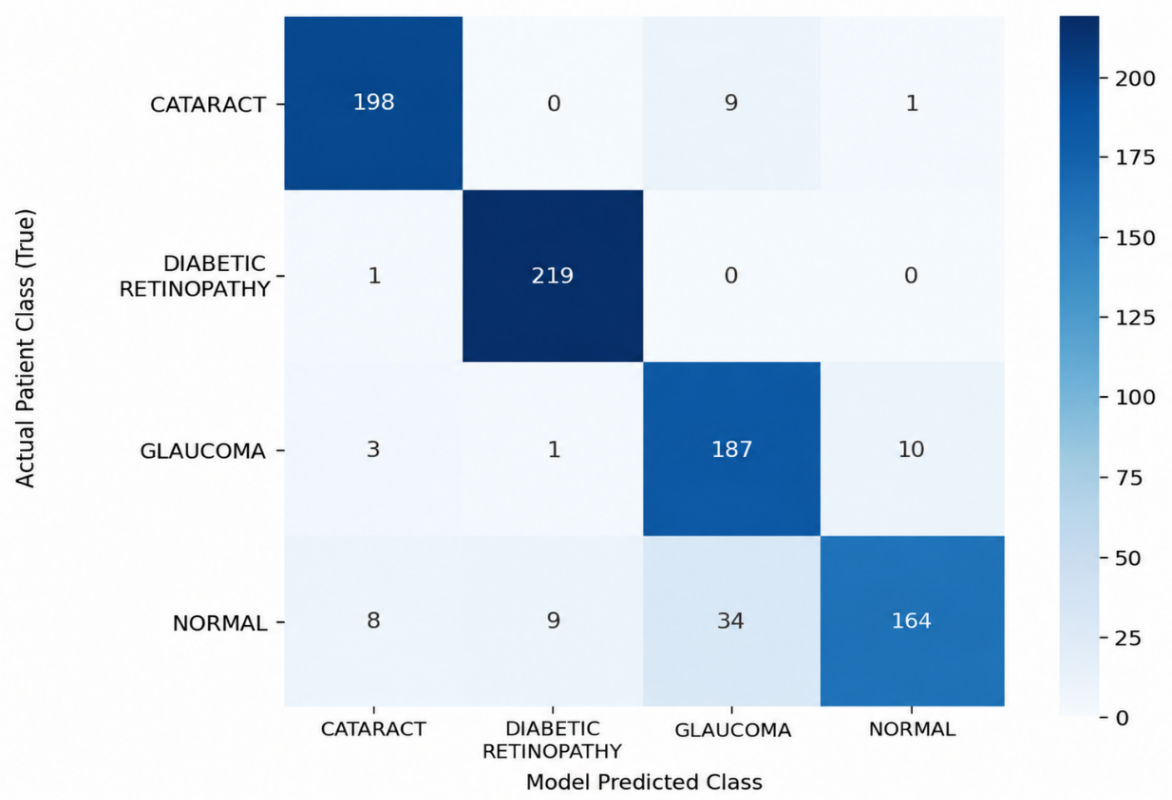


Figure 3. Confusion matrix MobileNetV3-Large after CLAHE preprocessing on data testing.

Confusion Matrix in the Form of Numbers

Table 2. Numerical confusion matrix of MobileNetV3-Large after CLAHE preprocessing on the test set

Actual\Prediction	Cataract	Diabetic Retinopathy	Glaucoma	Normal	Actual total
Cataract	198	0	9	1	208
Diabetic Retinopathy	1	219	0	0	220
Glaucoma	3	1	187	10	201
Normal	8	9	34	164	215
Total predictions	210	229	230	175	844

The main diagonal of the confusion matrix represents correct predictions: 198 + 219 + 187 + 164 = 768. The total number of incorrect predictions is 844 − 768 = 76. The accuracy calculated from this matrix is 768/844 = 0.90995, or 91.00%. This value falls within the approximate range of validation accuracy read from Figure 2 (approximately 89–93%). However, this agreement should be interpreted cautiously, because the validation values were read visually from the curve, test and validation accuracy are calculated from different subsets, and the evaluated model was the final epoch-30 model rather than a validation-selected model.

Metrics Analysis and Discussion

Overall Performance

Based on the confusion matrix (Table 3), the model achieves an accuracy of 91.00% (95% CI: 88.87–92.75%; Table 4). Macro-precision is 91.23%, macro-recall is 91.01%, and macro-F1 is 90.79% (95% CI: 88.80–92.66%). Macro-averaging assigns equal weight to each class, providing a more balanced assessment of performance across the four classes. The support-weighted F1-score is approximately 90.87%. The small difference between accuracy and macro-F1 suggests that the model achieves relatively strong overall performance, and the CI widths of approximately ±2 percentage points indicate reasonably precise estimates for a test set of 844 images; however, aggregate metrics alone do not reveal the types of errors that may be particularly important for individual classes.

Table 3. Aggregate performance metrics on the test set

Aggregate metric	Calculated value (95% CI)	Interpretation
Accuracy	91.00% (88.87–92.75)	The proportion of all correct test predictions
Macro-precision	91.23% (89.34–93.04)	Average precision of four classes of equal weight
Macro-recall	91.01% (89.16–92.80)	On average, four recalls of the same weight
Macro-F1	90.79% (88.80–92.66)	Average F1-score of four classes of equal weight

Weighted-F1	90.87% (88.83–92.79)	F1-score is averaged based on class support

Performance per Class

Table 4. Per-class performance metrics on the test set

Class	Precision (95% CI)	Recall (95% CI)	F1-score (95% CI)	Support
Cataract	94.29% (90.28–96.70)	95.19% (91.38–97.37)	94.74% (92.41–96.76)	208
Diabetic Retinopathy	95.63% (92.15–97.61)	99.55% (97.47–99.92)	97.55% (96.00–98.88)	220
Glaucoma	81.30% (75.77–85.81)	93.03% (88.65–95.81)	86.77% (83.08–90.02)	201
Normal	93.71% (89.10–96.45)	76.28% (70.16–81.47)	84.10% (80.10–87.80)	215

Note (Tables 3 and 4). Values in parentheses are 95% CIs, computed with the Wilson score method for accuracy, precision, and recall (Andersson, 2023) and with bootstrap resampling (10,000 resamples) for F1-scores and macro- and weighted-averaged metrics (El Jurdi et al., 2025).

The diabetic retinopathy class achieved the strongest overall performance, with a recall of 99.55%. Of the 220 images belonging to this class, 219 were correctly classified, while one was misclassified as cataract. These results indicate high sensitivity to diabetic retinopathy in the test subset. However, high recall on a single dataset does not guarantee comparable detection performance on images obtained from other sources. The cataract class achieved a precision of 94.29%, a recall of 95.19%, and an F1-score of 94.74%. The most frequent error involving this class was the misclassification of nine cataract images as glaucoma. Overall, the model performed relatively consistently on cataract images in the test set. Further evaluation is needed to determine whether these results remain consistent across different cameras, lighting conditions, and populations.

The glaucoma class achieved a recall of 93.03%, but its precision was lower, at 81.30%. This occurred because the model predicted 230 images as glaucoma, of which 187 were correctly classified. The most prominent source of false-positive predictions was the normal class, which contributed 34 images, followed by the cataract class, which contributed nine images. Thus, the model tended to classify some normal images as glaucoma. In screening applications, such errors could increase the number of false-positive referrals and should be investigated further through threshold analysis, probability calibration, and examination of relevant visual features.

The normal class had a recall of 76.28% (95% CI: 70.16–81.47%), the lowest among the four classes. Of the 215 normal images, 164 were correctly classified, while 34 were misclassified as glaucoma, nine as diabetic retinopathy, and eight as cataract. These results represent a major limitation of the experiment. Because the upper bound of this interval lies

below the lower bounds of the recall intervals for the other three classes, the weakness is unlikely to be explained by the sampling variability of the test set alone. Misclassification of normal images as glaucoma accounted for the largest number of cross-class errors. Systematic examination of misclassified images, label quality, and potential overlap between data sources is therefore important for understanding the underlying causes.

Interpretation of Error Patterns

Classification errors were not distributed evenly across the classes. The diabetic retinopathy class was recognized almost entirely correctly, whereas normal images were frequently misclassified as glaucoma. One possible explanation is that the visual features learned from certain normal images resemble patterns that the model associates with glaucoma. A similar difficulty was reported by Addyna et al., (2026), who identified glaucoma as the most challenging class and attributed this to its subtle morphological similarity to normal eyes. However, the confusion matrix alone cannot establish the biological or visual causes of these errors. Causal explanations would require examination of individual samples, expert validation, and interpretability methods such as Gradient-weighted Class Activation Mapping (Grad-CAM; Selvaraju et al., 2017), none of which were performed in this experiment.

Interclass confusion may also be influenced by labeling errors, dataset provenance, image quality, the distribution of image acquisition devices, and preprocessing choices. Therefore, error analysis should include a systematic collection of misclassified images and verification of their ground-truth labels. This study does not establish that the model can independently detect eye diseases in clinical practice; rather, it demonstrates its ability to classify the dataset labels with a certain level of performance.

The Role of CLAHE and the Claim Limits

Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied to the images before the dataset was split, and all reported experiments used this preprocessing configuration. Because CLAHE operates on each image independently, using only the pixel values of that image and fixed parameters, applying it before the split did not transfer information between the training, validation, and test subsets and therefore did not introduce data leakage. Under this experimental design, the observed accuracy of 91.00% can be described as the performance of the MobileNetV3-Large + CLAHE pipeline. However, this result does not establish that CLAHE improves accuracy relative to using the original images without contrast enhancement. To isolate the effect of CLAHE, a comparison group using images without CLAHE must be evaluated with the same data split, model architecture, random seed, data augmentation, optimizer, and checkpoint-selection procedure.

The literature supports empirical evaluation rather than the assumption that CLAHE invariably improves performance. Hayati et al. (2023) reported that CLAHE improved accuracy in some diabetic retinopathy classification models, whereas the ResNet34 model in their experiments performed better on the original images. Alwazzan et al. (2021) investigated fundus image enhancement using a combined approach involving CLAHE and Wiener filtering, which differs from the simpler pipeline used in this study. Differences in methods and datasets mean that previous findings should be treated as contextual evidence rather than as directly comparable benchmarks for the present experimental results.

Overfitting and Model Selection

The substantial difference between training accuracy, which approaches 99%, and validation accuracy, which remains approximately 90–93%, together with training loss approaching zero and validation loss increasing or fluctuating, provides strong evidence of possible overfitting. The model was trained for 30 epochs, and the model obtained after the final epoch was evaluated; neither best-checkpoint selection based on validation loss nor early stopping was applied. Consequently, the reported test score may not represent the model's best generalization performance and should be interpreted strictly as the performance of the final-epoch model.

Recommended improvements include saving the model checkpoint corresponding to the lowest validation loss, implementing early stopping, reviewing the learning curves, and repeating the experiment with multiple random seeds. Regularization strategies, such as appropriate data augmentation, label smoothing, or partial freezing of the pretrained backbone, may also be investigated in separate experiments. Any such modifications should be explicitly documented and should not be presented as part of the results obtained from the current experimental configuration.

Further Discussion and Implications

Comparison with Literature

The experiment achieved an accuracy of 91.00% and a macro-F1-score of 90.79% across four classes. These results are relevant to image classification research, but direct comparisons with other studies should be avoided when datasets, class definitions, data splits, and training procedures differ. Hayati et al. (2023) reported varying results across architectures and emphasized that the effect of CLAHE was not uniform. Kanjanasurat et al. (2023) also evaluated image enhancement techniques and several pretrained convolutional neural networks (CNNs) for diabetic retinopathy grading; however, their five-level severity classification task differed from the four-class classification task used in this study.

Agustini et al., (2026) reported a comparison of models for the same four fundus image categories and used both cross-validation and external validation. The similarity of the class categories makes their study a relevant contextual reference. Nevertheless, the results cannot be treated as a direct, head-to-head comparison because the experimental procedures, data splits, and models differ. Their study also highlights the possibility of reduced performance on external clinical images, underscoring the importance of cross-source evaluation.

Studies using comparable public four-class data provide further context. Alsohemi and Dardouri (2025) reported an accuracy of 95.12% for EfficientNetB3 on the same public dataset, Addyna et al., (2026) reported 89.17% for an optimized MobileNet on a 1,200-image Kaggle subset, and Agustini et al., (2026) reported a mean cross-validated accuracy of 0.85 for EfficientNet-B0. The accuracy obtained in this study (91.00%; 95% CI: 88.87–92.75%) therefore lies within or above the range reported for lightweight CNNs on similar data. Nevertheless, differences in data subsets, splitting strategies, preprocessing, and validation procedures, together with the absence of duplicate screening in this study, preclude any conclusion that the present configuration is superior.

Practical Implications

From a technical perspective, the model may serve as an initial component in research on image-based clinical decision support systems, particularly because recall for diabetic retinopathy and cataract was relatively high in the test set. However, the low recall for normal images and the relatively low precision for glaucoma indicate a risk of false-positive predictions and missed normal classifications that should not be overlooked. In healthcare settings, model errors may affect patient follow-up and clinical decision-making. Therefore, the model should not be used independently for diagnosis or as a replacement for assessment by qualified healthcare professionals.

Further development toward a clinical decision support system requires evaluation of sensitivity and specificity appropriate to the intended application, probability calibration, clinical review of representative samples, subgroup analysis, external validation, and post-deployment monitoring. Principles of health AI governance should also be considered, including transparency, privacy, accountability, and human oversight (World Health Organization, 2021).

Threats to Validity and Limitations

Table 5 summarizes the principal threats to the validity of this study. The non-CLAHE baseline, duplicate-image screening, correction of the drop_last setting, validation-based checkpoint selection, and repeated runs with multiple seeds were not performed in the current experiment and are therefore reported as limitations of the present results; external validation requires an independent labeled dataset, which was not available for this study.

Table 5. Threats to validity, potential impacts, and suggested mitigations

Threats	Potential impact	Suggested mitigation
There is no baseline without CLAHE	The effects of CLAHE cannot be isolated	Repeat with the original image and CLAHE on an identical split
Potential for image duplication	Test estimates can be overly optimistic	Exact/near-duplicate hash; split by patient if available
Resize ignores aspect ratio	Object shapes can be distorted	Compare resize with letterbox/padding
drop_last=True during training	Training metrics underestimated by at most 0.37 percentage points	Use drop_last=False or adjust the denominator
No best-checkpoint selection	The final model may be overfit	Select the checkpoint with minimum validation loss and apply early stopping
Single split and seed	Unmeasurable performance variation	Repeat multiple seeds or cross-validation
No external validation	Cross-source generalizations are not yet known	Test on independent datasets
Dataset metadata is incomplete	Limited reproducibility and interpretation	List sources, licenses, labels, and annotation procedures

Recommendations for Next Experiments

To answer the question of whether CLAHE improves classification performance, which the present study could not address, the next experiment should use an ablation design with at least two conditions: (A) original images without CLAHE and (B) images enhanced using CLAHE. Both conditions should use identical data splits, documented random seeds, the same model architecture, the same training-time data augmentation, and the same checkpoint-selection criteria. Additional comparisons could investigate different CLAHE parameters. However, changing one factor at a time is recommended to facilitate interpretation of the results.

To strengthen the report, macro-F1 should be retained as a primary evaluation metric, alongside a complete classification report and the numerical confusion matrix. Training and inference times should also be reported, and the test-set confidence intervals reported here should be complemented by variability estimates across multiple seeds or cross-validation folds. Misclassified samples should be examined systematically. If the ultimate goal is publication in a medical journal, a domain expert should review the labels and the interpretation of classification errors.

CONCLUSION

This study evaluated ImageNet-pretrained MobileNetV3-Large for the classification of four classes of eye images following CLAHE preprocessing. On the test set, the final-epoch model achieved an accuracy of 91.00% (95% confidence interval: 88.87–92.75%) and a macro-F1 of 90.79%. Recall was highest for diabetic retinopathy (99.55%) and lowest for normal images (76.28%); the principal weakness was the misclassification of normal images as glaucoma. The training and validation curves indicate possible overfitting.

Because the experiment did not include a baseline without CLAHE, the study cannot establish that CLAHE improves model performance; the conclusions are limited to a single experimental configuration on the internal dataset. The model should not be used independently for diagnosis. Future work should compare original and CLAHE-enhanced images, implement validation-based checkpoint selection, investigate duplicate images, and conduct external validation.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The author did not receive financial support for the research, authorship, or publication of this article.

CONFLICT OF INTEREST

The author declares no conflict of interest. This work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The author has no affiliation with the developers of the MobileNetV3-Large architecture or the CLAHE implementation used, or with the providers of the public dataset analyzed in this study.

AUTHOR CONTRIBUTIONS

Zahra Zhafirah: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing, and Project administration. The author has read and approved the final version of the manuscript and takes full responsibility for the integrity of the work.

GENERATIVE AI STATEMENT

During the preparation of this manuscript, the author used generative AI tools (including Claude, Anthropic) to improve language clarity and grammatical structure and to assist in locating, checking, and formatting bibliographic references. These tools were not used to generate or modify data, conduct the experiments, perform the statistical analyses, or interpret the results. All AI-assisted output was reviewed and edited by the author, every cited reference was verified against its original source, and the author takes full responsibility for the integrity and final content of this manuscript.

REFERENCES

Addyna, N., Sembiring, R. (2026). Enhancing Eye Disease Classification Accuracy Using Convolutional Neural Networks With Transfer Learning. *103.133.36.110*. Https://Doi.Org/10.33096/Ilkom.V18i1.2886.195-206

Agustini, L. Z., Bimantoro, F., & Dwiyansaputra, R. (2026). Eye Disease Classification Using Deep Learning: A Comparative Study Of Mobilenetv2, Xception, And Efficientnet-B0. *Jurnal Teknologi Informasi, Komputer, Dan Aplikasinya (Jtika)*, *8*(1), 1–13. Https://Doi.Org/10.29303/Jtika.V8i1.518

Alsohemi, R., & Dardouri, S. (2025). Fundus Image-Based Eye Disease Detection Using Efficientnetb3 Architecture. *Journal Of Imaging*, *11*(8), 279. Https://Doi.Org/10.3390/Jimaging11080279

Alwakid, G., Gouda, W., & Humayun, M. (2023). Deep Learning-Based Prediction Of Diabetic Retinopathy Using Clahe And Esrgan For Enhancement. *Healthcare*, *11*(6), 863. Https://Doi.Org/10.3390/Healthcare11060863

Alwazzan, M. J., Ismael, M. A., & Ahmed, A. N. (2021). A Hybrid Algorithm To Enhance Colour Retinal Fundus Images Using A Wiener Filter And Clahe. *Journal Of Digital Imaging*, *34*(3), 750–759. Https://Doi.Org/10.1007/S10278-021-00447-0

Andersson, P. G. (2023). The Wald Confidence Interval For A Binomial P As An Illuminating "Bad" Example. *The American Statistician*, *77*(4), 443–448. Https://Doi.Org/10.1080/00031305.2023.2183257

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). Imagenet: A Large-Scale Hierarchical Image Database. *2009 Ieee Conference On Computer Vision And Pattern Recognition*, 248–255. Https://Doi.Org/10.1109/Cvpr.2009.5206848

El Jurdi, R., Varoquaux, G., & Colliot, O. (2025). Confidence Intervals For Performance Estimates In Brain Mri Segmentation. *Medical Image Analysis*, *103*, 103565. Https://Doi.Org/10.1016/J.Media.2025.103565

Härtinger, P., & Steger, C. (2024). Adaptive Histogram Equalization In Constant Time. *Journal Of Real-Time Image Processing*, *21*(3), 93. Https://Doi.Org/10.1007/S11554-024-01465-1

Hayati, M., Muchtar, K., Roslidar, Maulina, N., Syamsuddin, I., Elwirehardja, G. N., & Pardamean, B. (2023). Impact Of Clahe-Based Image Enhancement For Diabetic Retinopathy Classification Through Deep Learning. *Procedia Computer Science*, *216*, 57–66. Https://Doi.Org/10.1016/J.Procs.2022.12.111

Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.-C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., & Le, Q. (2019). Searching For Mobilenetv3. *2019 Ieee/Cvf International Conference On Computer Vision (Iccv)*, 1314–1324. Https://Doi.Org/10.1109/Iccv.2019.00140

Huang, Z., Tang, C., Xu, M., & Lei, Z. (2020). Joint Retinex-Based Variational Model And Clahe-In-Cieluv For Enhancement Of Low-Quality Color Retinal Images. *Applied Optics*, *59*(28), 8628. Https://Doi.Org/10.1364/Ao.401792

Kanjanasurat, I., Anuwongpinit, T., & Purahong, B. (2023). Image Enhancement And 27 Pretrained Convolutional Neural Network Models For Diabetic Retinopathy Grading. *Sensors And Materials*, *35*(4), 1433. Https://Doi.Org/10.18494/Sam4084

Kapoor, S., & Narayanan, A. (2023). Leakage And The Reproducibility Crisis In Machine-Learning-Based Science. *Patterns*, *4*(9), 100804. Https://Doi.Org/10.1016/J.Patter.2023.100804

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key Challenges For Delivering Clinical Impact With Artificial Intelligence. *Bmc Medicine*, *17*(1), 195. Https://Doi.Org/10.1186/S12916-019-1426-2

Li, Z., Huang, J., Chen, J., Zeng, J., Jiang, H., Ding, L., Zhang, T., Sun, W., Lu, R., Zhang, Q., & Liang, L. (2024). A Deep-Learning Pipeline For The Diagnosis And Grading Of Common Blinding Ophthalmic Diseases Based On Lesion-Focused Classification Model. *Frontiers In Artificial Intelligence*, *7*, 1444136. Https://Doi.Org/10.3389/Frai.2024.1444136

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A. W. M., Van Ginneken, B., & Sánchez, C. I. (2017). A Survey On Deep Learning In Medical Image Analysis. *Medical Image Analysis*, *42*, 60–88. Https://Doi.Org/10.1016/J.Media.2017.07.005

Rahayu, T., Lestari, Y. D., Ayudianingrum, A., & Rif'ati, L. (2023). Spectacle Coverage Rate After Cataract Surgery In An Urban Area In Indonesia. *Clinical Optometry*, *15*, 167–173. Https://Doi.Org/10.2147/Opto.S417876

Reyad, M., Sarhan, A., & Arafa, M. (2023). A Modified Adam Algorithm For Deep Neural Network Optimization. *Neural Computing And Applications*, *35*, 17095–17112. Https://Doi.Org/10.1007/S00521-023-08568-Z

Sah, U. K., Chatterjee, J. M., & Sujatha, R. (2026). Multi-Class Eye Disease Classification Using Deep Learning Efficientnetb0 Fusion Techniques. *Scientific Reports*, *16*, 6368. Https://Doi.Org/10.1038/S41598-026-35357-0

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-Cam: Visual Explanations From Deep Networks Via Gradient-Based Localization. *2017 Ieee International Conference On Computer Vision (Iccv)*, 618–626. Https://Doi.Org/10.1109/Iccv.2017.74

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep Learning In Medical Image Analysis. *Annual Review Of Biomedical Engineering*, *19*(1), 221–248. Https://Doi.Org/10.1146/Annurev-

Bioeng-071516-044442

Sonali, Sahu, S., Singh, A. K., Ghrera, S. P., & Elhoseny, M. (2019). An Approach For De-Noising And Contrast Enhancement Of Retinal Fundus Image Using Clahe. *Optics & Laser Technology*, *110*, 87–98. Https://Doi.Org/10.1016/J.Optlastec.2018.06.061

Wanling, W., & Shah, N. M. (2025). Fundus Image Enhancement Using Clahe. *Journal Of New Explorations In Electrical Engineering*, *1*(1), 67–78. Https://Doi.Org/10.22452/Nece.Vol1no1.6

World Health Organization. (2021). *Ethics And Governance Of Artificial Intelligence For Health: Who Guidance*. World Health Organization.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How Transferable Are Features In Deep Neural Networks? *Advances In Neural Information Processing Systems*, *27*, 3320–3328. Https://Doi.Org/10.48550/Arxiv.1411.1792

Zhou, P., Xie, X., Lin, Z., & Yan, S. (2024). Towards Understanding Convergence And Generalization Of Adamw. *Ieee Transactions On Pattern Analysis And Machine Intelligence*, *46*(9), 6486–6493. Https://Doi.Org/10.1109/Tpami.2024.3382294