

XHARAC-Net: An Explainable Hybrid Attention Fusion and Residual Learning Framework for Automated Cataract Classification

Thanniru Pavan Vinayak¹ and Dr. Rajender Boini^{1*}

¹Research Scholar, Department of ECE, Chaitanya (Deemed to be University), Hyderabad, Telangana, India

^{1*}Professor, Department of EEE, Chaitanya (Deemed to be University), Hyderabad, Telangana, India

Email Id: ¹pavanvinayak.thanniru@chaitanya.edu.in and ^{1*}rajender_eee@chaitanya.edu.in

*Corresponding Author: rajender_eee@chaitanya.edu.in

Received: 14th June, 2026; Revised: 20th July 2026; Accepted: 20th August, 2026; Available Online: 05th September 2026

ABSTRACT

Cataracts are the leading cause of reversible blindness worldwide, accounting for nearly half of all cases of blindness. However, reliable automated grading of cataract progression stages remains an open and clinically consequential problem. The automated system currently lacks the ability to accurately assess cataract progression from the early stage to the advanced stage. The primary challenge arises from the subtle visual differences between normal, immature, and mature lens conditions. Deep learning models face difficulties because they use either complex ensemble systems that lose their useful features or single network designs that cannot identify essential details for precise-stage assessment. Therefore, in this study, we developed XHARAC-Net, which builds upon a residual convolutional network by adding squeeze-and-excitation (SE) and Convolutional Block Attention Module (CBAM) dual attention mechanisms to each residual stage of the network. The model uses attention mechanisms to enhance feature representations, which are then sent to a two-layer Artificial Neural Network (ANN) classifier that improves decision-making beyond the results achieved by the convolutional backbone. The design uses Gradient-weighted Class Activation Mapping (Grad-CAM) as an essential element that creates layer-wise saliency maps that demonstrate how the network develops from basic edge detection to specific lens opacity detection across different cataract development stages. The XHARAC-Net model achieved $99.92 \pm 0.13\%$ accuracy, $99.92 \pm 0.13\%$ F1-score, and $99.96 \pm 0.07\%$ specificity, enabling it to outperform all baseline methods that were reimplemented through a compact model that required only 2M parameters. The combination of strong discriminative performance, architectural simplicity, and built-in explainability positions XHARAC-Net as a practically deployable tool for AI-assisted ophthalmic screening in resource-constrained clinical settings in the future.

Keywords: Cataract, Convolutional neural networks, artificial neural networks, residual network, attention mechanism, Squeeze-and-Excitation, Feature interpretation, ocular images.

How to cite this article: Vinayak TP, Boini R, XHARAC-Net: An Explainable Hybrid Attention Fusion and Residual Learning Framework for Automated Cataract Classification. Int J Drug Deliv Technol. 2026; 16(7): 2071-2095. DOI: 10.25258/ijddt.16.7.215

Source of support: Nil.

Conflict of interest: None

1. INTRODUCTION

Cataracts are the main cause of blindness and one of the major causes of visual impairment worldwide. It is a result of the gradual opacity of the crystalline lens, which degrades the quality of the retinal image and visual function. Recent worldwide approximations reveal that cataracts are responsible for approximately 45–50% of blindness cases globally, thus, they cause more than 15–20 million people to be blind and approximately 80 million individuals to have moderate-to-severe vision impairment, especially among adults aged 50 years and older [1–3]. Even though effective surgical treatment is available, the absolute burden of cataracts continues to increase due to population aging and longer life expectancy, especially in low- and middle-income countries, where diagnostic coverage and timely intervention are still limited [2,3]. The figures above demonstrate cataracts as a clinical challenge and a significant public health problem; thus,

they call for the development of large-scale, objective, and accessible diagnostic strategies.

Cataract diagnosis involves the use of slit-lamp biomicroscopy together with a subjective evaluation system, such as the Lens Opacities Classification System III (LOCS III). Both techniques rely heavily on the expertise of the person conducting the examiner and involve significant variability among observers [4]. Although they are the gold standard for the clinical evaluation of cataracts, they are not practical for mass screening programs or teleophthalmology initiatives for the same reasons. Additionally, visual acuity measurements are insufficient for early detection, as patients may experience significant functional visual degradation, such as being bothered by glare, losing contrast, or having poor night vision, before any substantial loss of acuity becomes evident [5]. These shortcomings have led to a growing interest in imaging-

*Author for Correspondence: rajender_eee@chaitanya.edu.in

based diagnostics that can deliver objective, reproducible, and quantifiable evaluation of cataracts.

With the advancement of imaging techniques in ophthalmology, alternative diagnostic options are now available. Techniques such as Scheimpflug imaging and others that involve rotating cross-sectional imaging provide quantifiable information regarding the extent, density, and location of lens opacity and have been correlated with cataract severity and surgical planning in several studies [6,7]. However, the limitations of these techniques in terms of cost and infrastructure prevent their widespread use.

Fundus photography is another helpful technique. Although cataracts arise from the anterior segment of the eye, they leave distinct traces on retinal images, such as low contrast, blurring, light scattering, and nonuniform illumination [8,9]. The quality of a single fundus photograph can be rather noisy or erroneous, but looking at the same pattern of degradation in several images or an image sequence provides a more stable result, since temporal or multimodal data contribute to averaging out differences in illumination and the acquisition process. Thus, fundus image sequences can provide an indirect but sufficiently reliable method for detecting cataracts. Deep learning (DL) has accelerated this process. Sequence-based models based on ocular image sequences have provided cataract diagnosis accuracy comparable to that of an ophthalmologist [10–12] owing to the complementary nature of the information present in the sequence of frames and its contribution to the feature generalization ability. Given the considerable global burden of cataracts and the clear deficiencies associated with subjective diagnosis, the use of image sequences from the eyes is a reasonable path to pursue. The combination of this form of noninvasive and easy-to-obtain technology with computerized objective analysis can help advance this diagnostic method, as demonstrated in this study.

The remainder of this paper is organized as follows: Section 2 discusses existing approaches along with their pros, and cons. Section 3 describes the dataset and methodologies used to implement the proposed model. Section 4 presents the experimental outcomes and a comparison of the proposed framework with existing approaches. Finally, Section 5 presents the conclusions.

2. RELATED WORKS

This section discusses several recently introduced machine learning (ML) and deep learning (DL) approaches for cataract detection from retinal or ocular images.

Agustin et al. [13] developed a convolutional neural network (CNN)-based cataract detection system. Initially, they split the given eye image into R, G, and B components and then employed the Canny edge to identify significant edges. Finally, a customized CNN framework was applied to categorize cataract and normal eye images. Lahari et al. [14] presented a novel cataract state detection

network (CSDNet)- based on retinal fundus images. The system includes data preparation, augmentation, and preprocessing, followed by a CNN. Nguyen et al. [15] suggested a hybrid CNN framework using the sequence of operations such as contrast enhancement by contrast-limited adaptive histogram equalization (CLAHE), data augmentation using geometric transformations, feature extraction by the integration of baseline DensNet121 with fine tune DensNet121 architecture and classification by the ensemble learning. Sariera et al. [16] implemented an automated system for the detection of cataracts using CLAHE, grey-level co-occurrence matrix (GLCM), first-order statistics (FOS), statistical feature matrix (SFM), neighborhood gray tone difference matrix (NGTDM), and random forest (RF).

Mahmood et al. [17] developed an enhanced ResNet50 architecture with transfer learning for the identification of cataracts through retinal images. Khan et al. [18] also introduced ResNet50 to discriminate different stages of cataracts such as normal, moderate and severe in retinal images. Santoso et al. [19] analyzed various CNN models, such as ResNet-50, EfficientNet-B1, and VGG-16, to predict cataracts from retinal images. Among these three models, the ResNet50 exhibited superior performance. Ramdhani et al. [20] proposed a customized CNN architecture for the early diagnosis of cataracts from retinal images. In addition, they tested their proposed CNN model using various optimization methods, including Adaptive Moment Estimation (Adam), Root Mean Square Propagation (RMSprop), Adaptive Gradient Algorithm (AdaGrad), and Stochastic Gradient Descent (SGD). The results showed that the CNN model with the Adam optimizer performed the best.

Ramakrishnan et al. [21] proposed a hybrid deep learning architecture by combining the strengths of MobileNet, ResNet50, Inception V3, and VGG19, and finally classified the images into cataractous and normal retinal images using a Support Vector Machine (SVM) classifier. Syahda et al. [22] evaluated different deep learning and ensemble models, namely MobileNet, VGG16, VGG16+-MobileNet (ensemble model), MobileNet-SVM, and VGG16-SVM. Among these, the ensemble model performed the best. Islam et al. [23] designed a new automated cataract detection system using seven traditional deep learning models, followed by Principal Component Analysis (PCA) and different supervised machine learning classifiers. Among these combinations, MobileNetV2-PCA-RF achieved a high classification accuracy. Fahmi and Estiko

[24] suggested a traditional residual network mechanism, such as ResNet50, and ResNet101 for the prediction of cataracts. Among these two models, ResNet50 yielded a better accuracy. Kaushik and Sharma [25] introduced a modified ResNet50 architecture with transfer learning for automated ophthalmic diagnosis.

For better interpretation, all the above-mentioned issues are tabulated in Table 1 with their methodology, pros, and cons.

Table 1: Summary of the state-of-the-art methods

Reference	Methods	Pros	Cons
Agustin et al. [13]	Canny edge detection + Customized CNN	Edge-based preprocessing highlights lens opacity boundaries, and customized CNN reduces architectural complexity and training cost.	Canny edge detection is sensitive to noise and illumination variations, and RGB splitting may ignore deeper texture patterns relevant to cataracts.
Lahari et al. [14]	CSDNet	The end-to-end DL framework improves robustness, and augmentation helps mitigate the limited dataset size.	The performance of these models heavily depends on the quality of fundus images and lacks explicit explainability of learned features.
Nguyen et al. [15]	CLAHE + geometric augmentation + DenseNet121 (fine-tuned) + ensemble learning	CLAHE enhances low-contrast cataract regions, DenseNet improves feature reuse, and ensemble boosts classification accuracy.	High computational complexity: Ensemble learning increases inference time and memory usage.
Al Sariera et al. [16]	CLAHE + handcrafted features (GLCM, FOS, SFM, NGTDM) + Random Forest	Interpretable handcrafted features: RF handles feature redundancy well.	Limited representation power compared to DL; performance degrades on large-scale- or diverse datasets.
Mahmood et al. [17]	Transfer learning with enhanced ResNet50	Deep residual learning captures complex cataract patterns, and transfer learning reduces the training data requirements.	Risk of overfitting on small datasets; ResNet50 is computationally intensive.
Khan et al. [18]	ResNet50	It can distinguish cataract severity levels and has strong feature extraction via residual blocks.	It requires large annotated datasets, and class imbalance can affect stage-wise accuracy.
Santoso et al. [19]	ResNet50, EfficientNet-B1, VGG16	Comprehensive benchmarking showed that ResNet50 exhibited superior performance.	Does not address optimization or hybridization; limited discussion on real-time-deployment.
Ramdhani et al. [20]	Customized CNN	Optimizer analysis provides training insights, and Adam improves convergence stability.	Custom CNN may lack generalization and have limited depth compared to modern architectures.
Ramakrishnan et al. [21]	Hybrid DL (MobileNet, ResNet50, InceptionV3, VGG19) + SVM	It combines multi-scale and deep semantic features, and SVM improves the decision boundary.	Feature fusion increases complexity, and training and tuning are computationally expensive.
Syahda et al. [22]	Pretrained CNNs + ensemble	The ensemble improves the robustness and accuracy, and the comparative evaluation strengthens the conclusions.	Increased inference time: Ensemble models are harder to deploy in low-resource- settings.
Islam et al. [23]	Multiple DL models + PCA + RF	PCA reduces dimensionality, and RF improves classification stability.	PCA may discard discriminative deep features, and the pipeline complexity is high.
Fahmi & Estiko	ResNet50 and	Deep residual learning is	Very deep networks require high

[24]	ResNet101	effective for cataract patterns, and ResNet50 balances depth and accuracy.	computational resources and have limited interpretability.
Kaushik & Sharma [25]	Modified ResNet50 with transfer learning	Architectural modifications improve feature adaptability and are effective for automated diagnosis.	Dataset dependency: The performance of the model may drop across different imaging devices.

Highlights of the study:

To address the issues in the existing models, as listed in Table 1, we introduced a hybrid architecture, namely XHARAC-Net, by integrating a convolution block attention mechanism (CBAM), a squeeze-and-excitation (SE) mechanism, residual learning with attention, artificial neural networks (ANN), and gradient-weighted class activation mapping (Grad-CAM). The main highlights of the proposed model are as follows:

1. We suggest a novel residual convolutional structure in which channel and spatial attention mechanisms (squeeze-and-excitation and CBAM) are integrated into every single residual block of the neural network instead of being post-processed after the feature extraction stage, as in most current cataract grading models. Such a structure can recalibrate the importance of features on a progressively lower- and higher-level abstraction scale and will be beneficial in separating immature and mature cataracts, which are very similar and differ from each other only in terms of subtle changes in texture and opacities.
2. We demonstrate that the proposed stage-wise attention mechanism results in a very high discrimination accuracy rate but with a significantly reduced number of parameters compared to deep and ensemble-based models used in previous studies on cataract grading. This proves that well-applied attention can be a much more efficient method than scaling or ensembles in cataract classification tasks.

3. We propose a two-step classification pipeline, where the deep residual features are passed onto an artificial neural network classifier, which performs the final task of making decisions and explicitly quantifies the effect of the second step using matching ablation tests. The results show that separating feature extraction from classification is beneficial in terms of improving accuracy and stability, rather than assuming that the addition of a classification head will improve the system performance.
4. Saliency-based interpretability analysis is extended not only to one terminal layer but to multiple depths of the network, and we find that there is a clear transition from low-level edge activation to opacity-specific clinical activation as we increase the network depth, thus providing interpretability grounded on the underlying disease process and not on one-time visualization.

All results are presented after running experiments using both hold-out evaluation and multiple fold cross-validation, and the variability is measured across multiple folds and not in a single train/test split, as done in previous studies.

3. THE PROPOSED METHODOLOGY

This section describes the dataset and methods used to implement the proposed method. Figure 1 illustrates the workflow of the proposed model, including data collection, feature engineering using CNN, and classification using ANN.

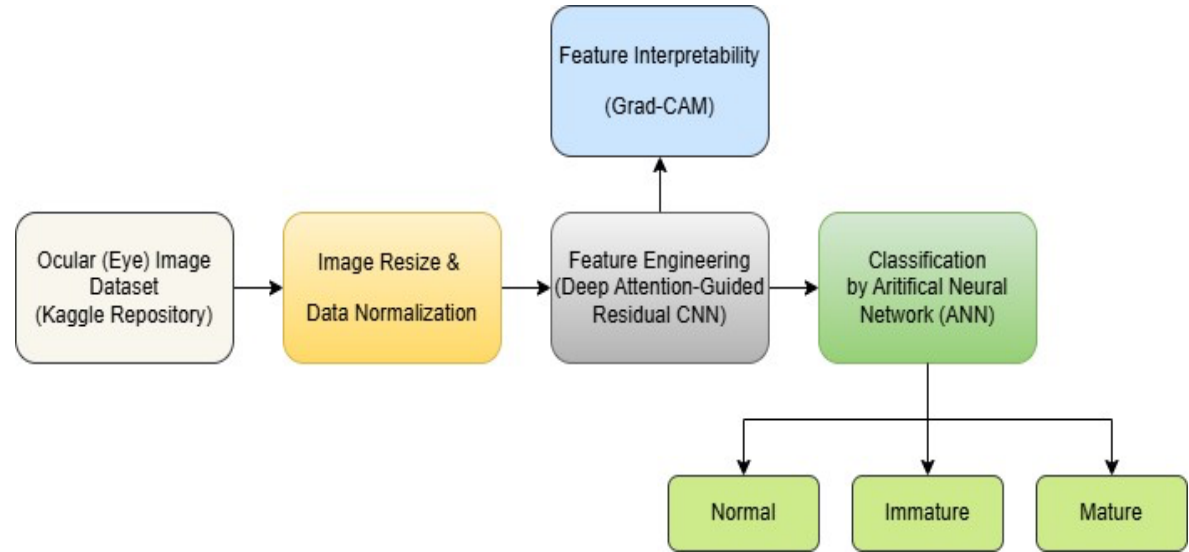


Figure 1: Block diagram of the proposed XHARAC-Net system

3.1. Data Collection

To validate the proposed and existing models, we collected 6,186 ocular (eye) cataract data samples from the open-access repository Kaggle [26]. The dataset included three classes: 2,166 normal eyes without cataracts, 1,880 immature cataract eyes in which lens opacity was just beginning, and 2,140 mature cataract eyes in which cloudiness was extensive and vision impairment was severe. By delineating the visual characteristics of healthy

eyes and different stages of cataract development, the dataset aids in the creation of automated systems that would support ophthalmologists in the areas of early diagnosis, disease monitoring, and treatment planning. Figure 2 shows sample eye cataract images. Before performing feature engineering, all images were resized to 256×256 dimensions and then normalized using the Z-score to improve the network speed.

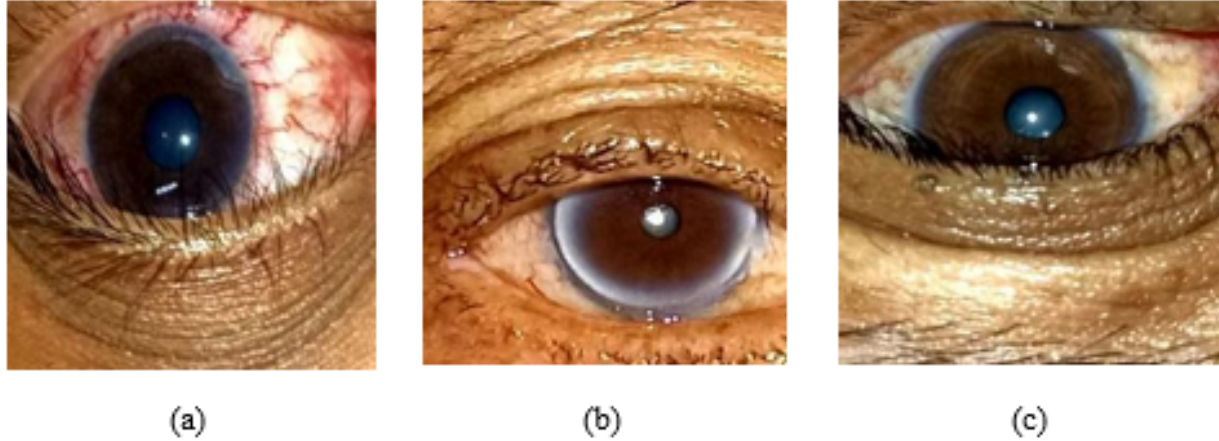


Figure 2: Sample eye cataract images: (a) Normal; (b) Mature; (c) Immature

3.2. Feature Engineering

Automated detection of eye cataracts involves the use of a feature extraction procedure that can detect changes in texture degradation, contrast loss, and opacity of the image at various levels of cataracts. Normal convolutional neural networks (CNN) find it challenging to distinguish between immature and mature cataracts owing to the similarities within the class and low contrast between the classes in eye images. To solve these challenges, this study introduces the use of a novel deep attention-guided residual CNN (DAG-ResCNN) that uses hierarchical feature learning and adaptive feature recalibration.

The neural network learns to extract discriminative features directly from grayscale eye images in a full cycle, while stabilizing gradient propagation and avoiding overfitting. Combining the effects of residual networks with channel and spatial attention enables the proposed network architecture to focus on important clinical information at multiple abstraction levels. Therefore, this approach enables the modeling of local texture changes create complementary channel descriptors according to Eq. (1) and (2), respectively

$$z_{avg}^c = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W X_c(i, j) \quad (1)$$

where, X_c refers to the c^{th} channel of the input feature map.

$$z_{max}^c = \max_{\{i,j\}} X_c(i, j) \quad (2)$$

These descriptors are then subjected to shared fully connected layers with a reduction ratio r of 8, following

and global opacity patterns. Consequently, the performance of the proposed model in the classification of normal, immature, and mature cataracts was enhanced. In the following subsections, we discuss the blocks used in the implementation of the proposed feature-engineering architecture.

A. Convolutional Block Attention (CBAM)

The Convolutional Block Attention Module (CBAM) is used to enhance feature representation by relying on the sequential application of channel attention and spatial attention mechanisms. An intermediate feature map $X \in \mathbb{R}^{H \times W \times C}$, is taken as an input, where

H and W are the height and width of the input feature map, respectively, and C indicates the number of channels in the feature map. The CBAM initially determines which feature channels are crucial, and then locates the informative areas. In the channel attention submodule, both global average and max pooling are performed across spatial dimensions to

sigmoid activation to yield channel attention weights. This process can be demonstrated using Eq. (3) as follows:

$$M_c = \sigma \left(MLP(z_{avg}) + MLP(z_{max}) \right) \quad (3)$$

Here, $MLP(\cdot)$ indicates the shared multilayer perceptron in the CBAM channel attention module, and $\sigma(\cdot)$ denotes the sigmoid activation function.

The input feature map was reassessed using channel-wise multiplication. Subsequently, spatial attention is applied by computing the average and maximum projections along the channel axis and concatenating them using Eq. (4):

$$M_s = \sigma \left(f^{(7 \times 7)}([AvgPool(X); MaxPool(X)]) \right) \quad (4)$$

where, $[;]$ indicates the channel-wise concatenation operation and $f^{(7 \times 7)}(\cdot)$ denotes the convolution operation with a 7×7 kernel.

The CBAM-refined output was obtained through element-wise multiplication of the feature maps. The entire CBAM process is shown in Figure 3.

The resulting spatial attention map delineated cataract-affected areas, such as lens opacity and texture

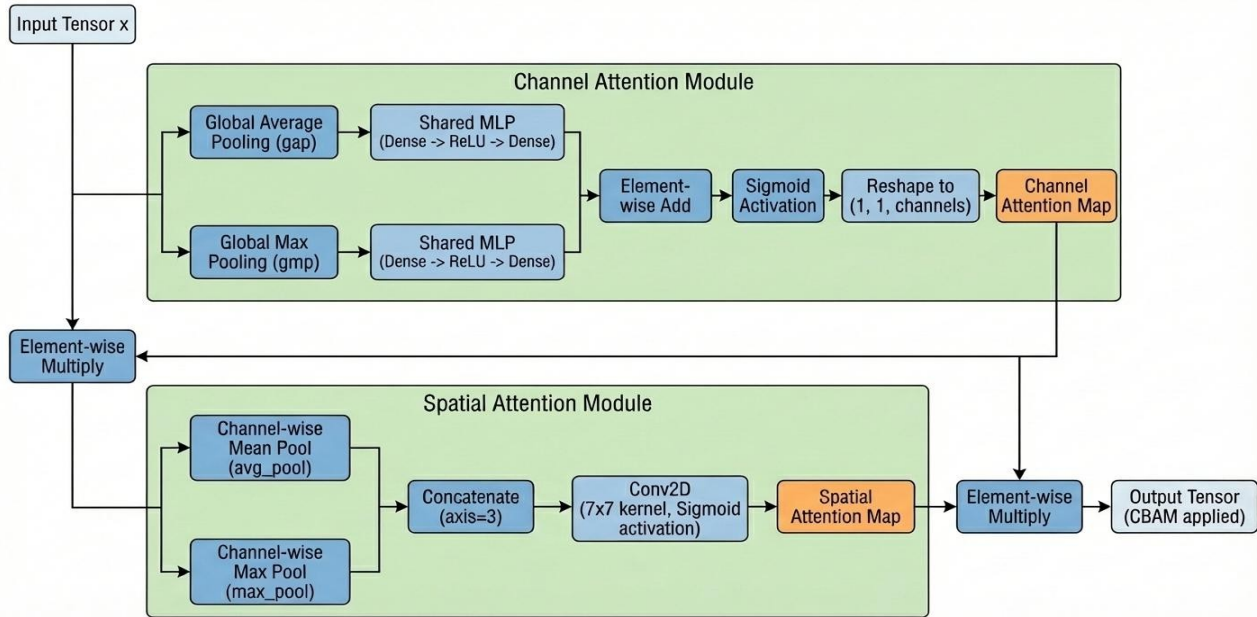


Figure 3: The Proposed CBAM architecture

B. Squeeze-and-Excitation (SE) Attention Module

The Squeeze-and-Excitation (SE) module shown in Figure 4 acts as a minor attention mechanism on channels to represent their relationships. For any feature map X , the squeeze step first gets the average of each channel over the

$$s_c = \sigma \left(W_2 \delta \left(W_1 z_{avg}^c \right) \right) \quad (5)$$

In this equation, W_1, W_2 are the learnable weight matrices corresponding to the two fully connected layers in the SE

entire area as in eq. (1). The intermediate descriptor is then fed through two fully connected layers with ReLU and sigmoid activations to obtain the channel scaling factors, s_c , as per the eq. (5):

$$\tilde{X}_c = s_c \cdot X_c \quad (6)$$

This process allows the network to call attention to features that are very informative with respect to cataracts and then to muffle the channels that are not so important,

block. Subsequently, the original feature map is calibrated again using Eq. (6):

thus improving the separation of normal, immature, and mature cataracts.

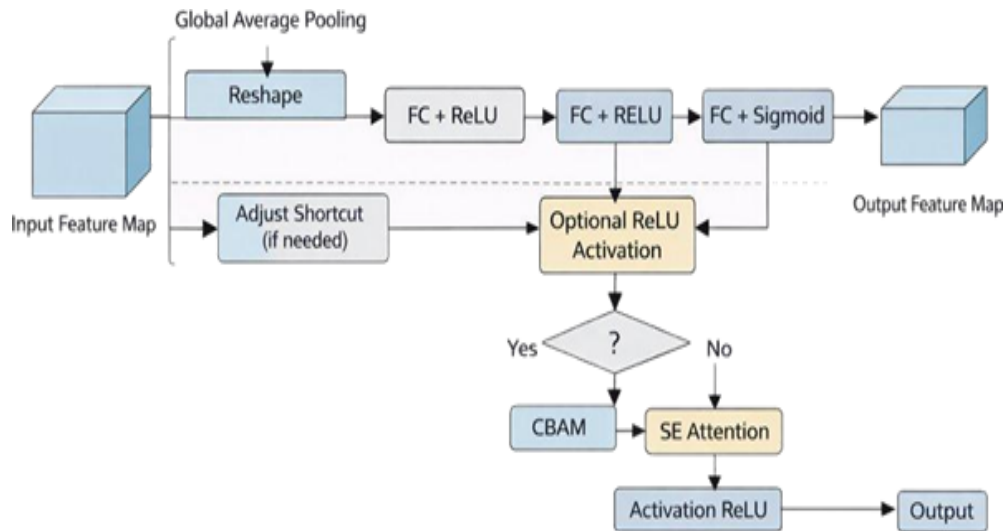


Figure 4: The SE architecture

C. Residual Block with Attention

The residual block is purposely built to support the extraction of deep features and simultaneously address the vanishing gradient problem. The structure contains two

$$Y = F(X) + X$$

where Y indicates the output feature map, and $F(X)$ depicts the residual mapping learned by the stacked convolutional layers. In the case of different numbers of channels, a 1×1 convolution was utilized in the shortcut path to conveniently adjust the dimensions.

The attention mechanism is applied after residual fusion to

layers of 3×3 convolutions along with Batch Normalization and ReLU activation in that order. The residual learning formulation is expressed as follows:

$$(7)$$

enhance the discriminative features (SE or CBAM). The combination of residual learning and attention improves the robustness of the model against noise and makes it capable of detecting subtle cataract progression characteristics by learning. The entire residual mechanism process, along with the attention method, is illustrated in Figure 5.

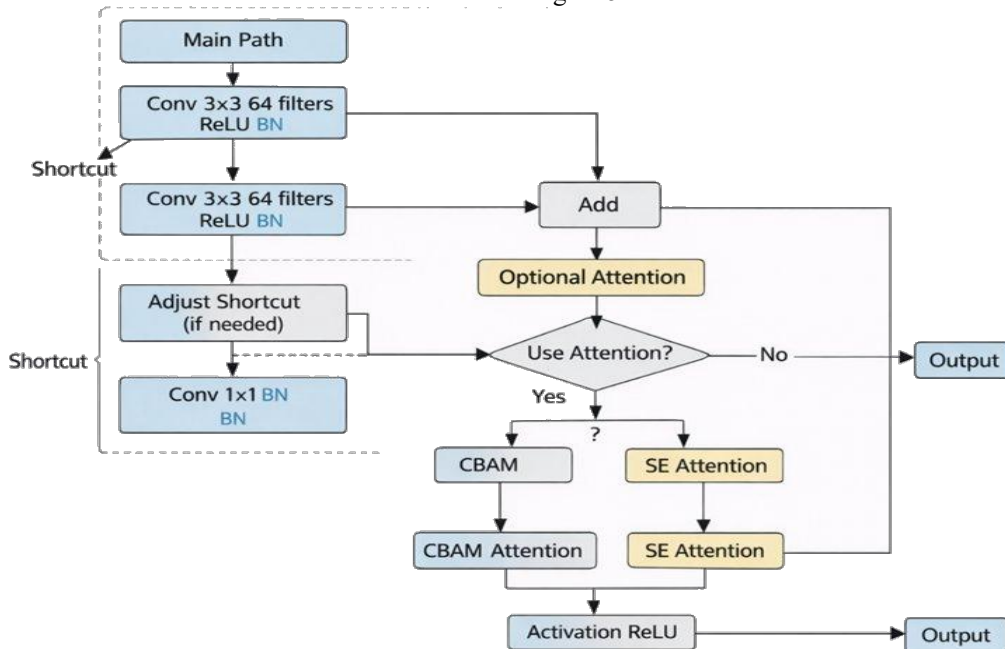


Figure 5: Architecture of the residual block with attention mechanism

D. Overall Architecture

The proposed deep learning model is a residual convolutional neural network model that employs channel attention and is designed particularly for cataract image classification. A deep learning model was proposed to

accomplish the task based on grayscale input images of size $256 \times 256 \times 1$. Therefore, the network can efficiently learn the texture degradation and opacity changes associated with cataract progression. The input layer of the network accepts grayscale eye image inputs. Although the

network receives less complex input because of its grayscale nature, the intensity information of clinical importance is not lost.

The First stage concentrates on detecting low-level visual patterns, such as edges, contrast loss, and signs of opacity. A 3×3 convolution with 64 filters was applied, and then Batch Normalization was performed to stabilize the training process. The output of this layer is maintained as a shortcut branch. Two additional 3×3 convolutional layers were used to further enhance the feature representation. A max-pooling layer using a 2×2 window reduces the spatial resolution from 256×256 to 128×128 ; thus, the receptive field increases. To maintain the consistency of the residual learning through the downsampling process, the shortcut branch is forwarded through a 1×1 convolution with stride 2 and Batch Normalization. The main and shortcut paths are then joined through element-wise addition, resulting in the formation of residual blocks. Subsequently, the SE attention module is applied after residual fusion to recalibrate the channel-wise features, allowing the network to focus more on early cataract-relevant structures.

During the second stage, the number of filters was increased to 128 to capture more complicated patterns, such as the distribution of lens opacity and structural degradation. Similar to the first stage, the first convolution is followed by two more convolutional layers, each with Batch Normalization and ReLU activation. The spatial resolution was further reduced to 64×64 using max-pooling. A residual shortcut connection with a 1×1 convolution and a stride of 2 guarantees the alignment of the dimensions before the element-wise addition is performed. After the residual operation, the SE attention module was applied again, allowing for adaptive emphasis on the discriminative mid-level features that differentiate

between immature and mature cataracts.

In the third stage, the representation of high-level semantics related to the appearance of severe cataracts was obtained. In this case, the depth of the features was 256 channels, and the dimensions of the image were 32×32 . Again, the same architecture was used: the convolutional layers were stacked using Batch Normalization, max-pooling to down sample, and a residual connection for proper gradient flow. The introduction of the SE attention at this stage allows the network to properly increase the importance of high-level feature channels that are important for cataract classification.

In the final convolutional layer, the Global Average Pooling (GAP) process is performed, where the spatial dimension is reduced to one feature vector. By performing such action, it reduces the number of trainable parameters in comparison to flattening process while improving the generalization through preserving global spatial consistency in feature learning. The feature vector obtained from the pooling process was fed into the classification process using several neural networks. Initially, the fully connected (FC) layer with 120 neurons and ReLU activation is followed by Batch Normalization and dropout processes that are used to solve the problem of overfitting. The second FC layer containing 50 neurons continues to work with the input data, and the dropout process is applied again for regularization. The output layer contains three neurons with a softmax activation function that produces probability scores for the three cataract classes.

The Global Average Pooling (GAP) operation compresses the spatial feature map of dimensions $H \times W \times C$ into a single vector of length C by computing the spatial mean across all positions, as expressed in Eq. (8):

$$GAP(X) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W X(i, j) \quad (8)$$

where $X(i, j)$ is the activation value at spatial position (i, j) . The resulting feature vector is subsequently passed through an ReLU-activated fully connected layer. The

final classification scores for each cataract class c are computed using the softmax function, as expressed in Eq. (9):

$$P(y = c | x) = \frac{\exp(z_c)}{\sum_{k=1}^C \exp(z_k)} \quad (9)$$

where z_c is the logit score for class c , $C = 3$ is the total number of cataract classes (Normal, Immature, Mature), and x is the input feature vector. The softmax output directly yields a probabilistic prediction used during inference.

The complete system employs a combination of deep residual learning, feature scaling, and attention to the channel to distinguish minor visual differences between normal, immature, and mature cataract images. The judicious positioning of squeeze-and-excitation (SE) attention after every residual fusion guarantees that the clinically important features are continuously highlighted across the network, thus paving the way for a powerful

and flexible system for automatic cataract diagnosis.

3.3. Classification

To improve the performance of the proposed model, further features from the dense layer with 50 units of the suggested feature engineering model, DAG-ResCNN, are fed to an artificial neural network (ANN), namely HARAC-Net, to discriminate cataracts from ocular (eye) images. Finally, the performance of the ANN was analyzed using well-received metrics, including specificity, recall, precision, F1-score, and accuracy [27].

The evaluation metrics used in this study were derived from the confusion matrix entries: True Positives (TP), True Negatives (TN), False Positives (FP), and False

Negatives (FN). These metrics are formally defined as follows: (10)–(14).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

$$Specificity = \frac{TN}{TN + FP} \quad (13)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (14)$$

For the multi-class evaluation, the macro-averaged form of each metric was computed by averaging the per-class metric values uniformly across all $C = 3$ cataract classes (Normal, Immature, Mature), thereby giving equal weight to each class regardless of its sample count.

4. RESULTS AND DISCUSSIONS

In this section, we discuss the experimental results that demonstrate the efficiency of the proposed approach and compare it with other recently developed approaches presented in Section 2. To investigate the significance of the proposed algorithm in the prediction of cataract diseases, extensive simulations were performed based on two data conditions, namely, Hold-out, and K-fold cross-validation (K-FCV).

K-FCV is a popular and computationally effective method

for preventing overfitting. The selection of K plays an extremely important role in classification problems because lower K values typically result in higher bias and lower variance; consequently, the generated model is prone to underfitting. In contrast, higher K values would yield less bias and high variance, causing potential overfitting. Hence, to achieve a fair balance between bias and variance, a moderate K value of 5 was used in our study. In addition to K-FCV, the hold-out validation technique was applied, wherein 80% of the data were trained and 20% were tested.

The expected generalization error of a model can be decomposed into bias and variance components, as expressed in Eq. (15), which motivates the use of K-fold cross-validation to balance this trade-off:

$$E((f(x) - y)^2) = Bias^2(f(x)) + Var(f(x)) + \sigma^2 \quad (15)$$

where σ^2 is the irreducible noise. The K-fold cross-validation average performance over K folds is computed

as follows: (16), providing a statistically reliable estimate of the generalization performance:

$$CV(K) = \frac{1}{K} \sum_{k=1}^K ACC_k \quad (16)$$

where ACC_k is the classification accuracy on the k -th held-out fold, and $K = 5$ in our experiments.

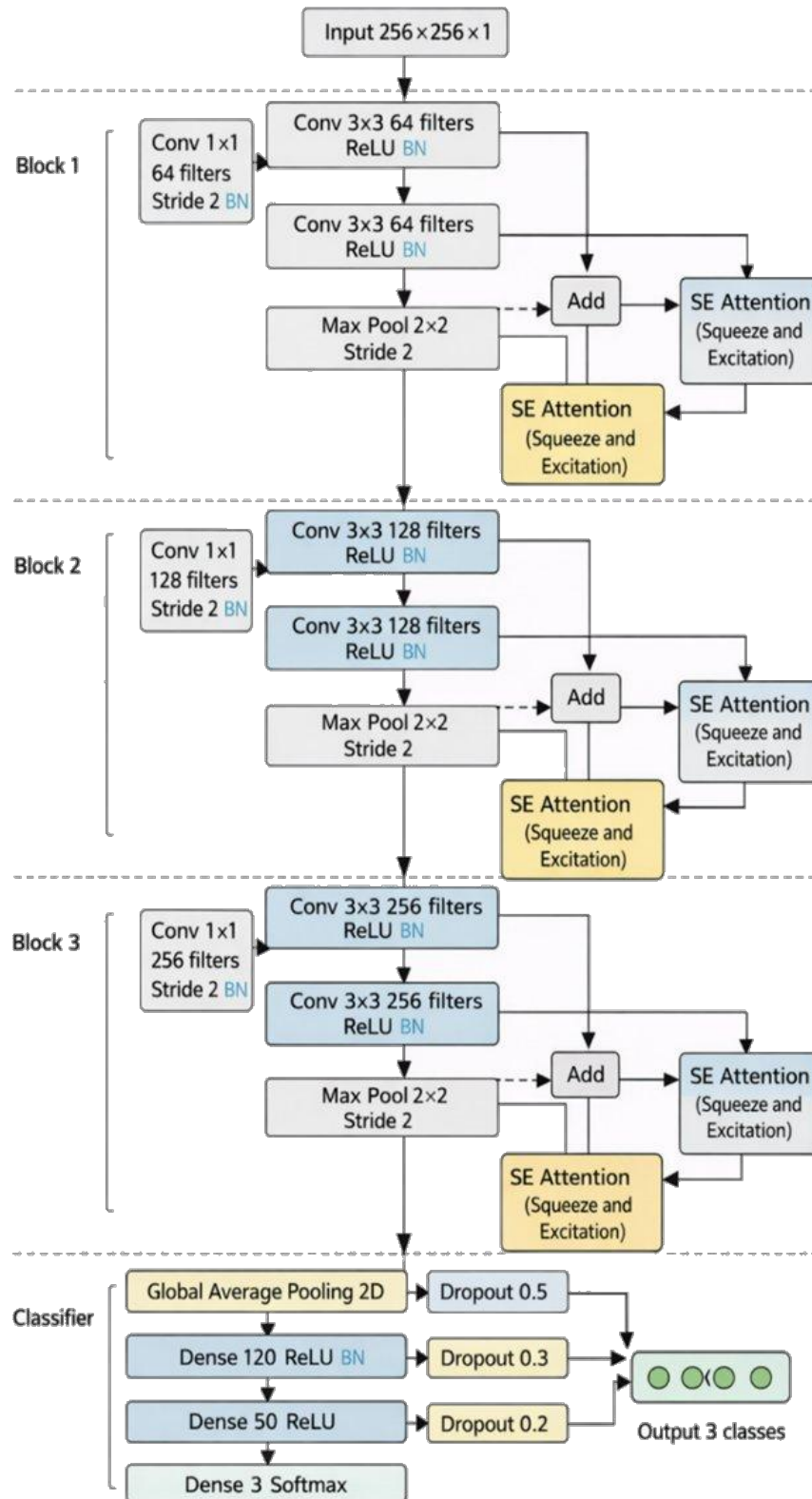


Figure 6: Architecture of the suggested DAG-ResCNN feature engineering network

4.1. Training Configurations

This subsection discusses the configurations of the training parameters used to train the proposed CNN and ANN models.

To effectively categorize cataract samples from ocular (eye) images, we employed a DAG-ResCNN framework to

extract significant features. Typically, our architecture automatically attempts to attain the relevant features using a series of hidden layers and learns using the back-propagation approach. During the training process, the predicted and actual class distributions were compared using a categorical cross-entropy loss function, which

indicated the extent of the discrepancies. Model optimization was performed using the Adam optimizer, which automatically adjusted the learning rates according to the training requirements. A learning rate of 0.0001 was selected to ensure stable convergence. The network was trained for 50 epochs, and the batch size was set to 32 to control the speed of convergence and the performance of

$$L = -\frac{1}{N} \sum_{n=1}^N \sum_{c=1}^C y_{n,c} \log P(y = c | x_n) \quad (17)$$

where N is the total number of training samples, $C = 3$ is the number of cataract classes, $y_{\{n,c\}}$ is the ground-truth one-hot label (1 if sample n belongs to class c , else 0), and $P(y = c|x_n)$ is the predicted softmax probability for sample

$$m_t = \beta_1 m_{\{t-1\}} + (1 - \beta_1) \nabla L(\theta_{\{t-1\}}) \quad (18)$$

$$v_t = \beta_2 v_{\{t-1\}} + (1 - \beta_2) [\nabla L(\theta_{\{t-1\}})]^2 \quad (19)$$

$$\theta_t = \theta_{\{t-1\}} - \eta \times \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}} \quad (20)$$

where m_t and v_t are the first and second moment estimates at step t , $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ are hyperparameters, $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected estimates, $\eta = 0.0001$ is the learning rate, and θ represents the model parameters.

To improve the performance of the proposed model, we employed an ANN classifier. The ANN layout contained two hidden layers that were completely connected and composed of 150 and 50 neurons, respectively. The ReLU activation function was applied to the hidden layers to

$$ReLU(z) = \max(0, z) \quad (21)$$

The L2 regularization term incorporated in the ANN loss penalizes large weight magnitudes to prevent overfitting,

$$L_{total} = L + \lambda \sum_i w_i^2 \quad (22)$$

where $\lambda = 0.0001$ is the regularization coefficient, w_i are the network weights, and L is the categorical cross-entropy defined in Eq. (10). This penalty discourages the weight

generalization. The training parameters of the CNN are presented in Table 2.

The categorical cross-entropy loss function used to train the DAG-ResCNN model is defined in Eq. (17), which measures the divergence between the predicted probability distribution and the true one-hot-encoded label:

n and class c . The Adam optimizer updates the model parameters using adaptive first- and second-order moment estimates, as defined in Eqs. (18)–(20):

introduce non-linearity and enhance the learning process. The network was optimized using the Adam optimizer with a maximum training limit of 500 iterations. To combat overfitting, L2 regularization with a coefficient of 0.0001 was used. The complete training configuration of the ANN model is listed in Table 3.

The ReLU activation function, applied to the hidden layers of the ANN, introduces non-linearity while maintaining computational simplicity and is defined in Eq. (21):

extending the total training objective to Eq. (22):

magnitudes from growing excessively large, thereby improving the generalization to unseen cataract images.

Table 2: Training parameters of the proposed CNN architecture

Parameter	Value
Loss function	Categorical Cross-Entropy
Optimizer	Adam
Learning rate	0.0001
Epochs	50
Batch size	32

Table 3: Training parameters of the ANN

Parameter	Value
Hidden layers	2
Neurons per layer	150, 50
Activation function	ReLU
Optimizer	Adam
Maximum iterations	500

Regularization (L2)	0.0001
---------------------	--------

4.2. Hold-out Analysis of the Proposed Model

In this section, we evaluate the performance of the suggested model using the hold-out validation process. In this study, 80% of the data were utilized for training purposes, while the remaining 20% were used for testing, thus creating an optimal balance between learning and performance analysis.

A. Quantitative Analysis

The proposed deep attention-guided residual CNN (DAG-ResCNN) and the hybrid model of DAG-ResCNN-ANN (HARAC-Net) were objectively analyzed using the hold-out validation approach for cataract image classification, as shown in Table 4. These models were analyzed in terms of three clinically important classes: Normal, Immature and Mature, using the evaluation criteria provided in Section 3.3.

As depicted in Table 4, both DAG-ResCNN and HARAC-Net achieved excellent performance with a percentage of 100% for all metrics in the classes, indicating that the two types of cataracts (normal and various stages) were

successfully discriminated from each other. This performance consistency for all classes confirms the reliability of the proposed framework for feature extraction and classification using the hold-out approach.

The confusion matrices for (a) DAG-ResCNN and (b) HARAC-Net are presented in Figure7. Confusion matrices showed that all tested samples were properly classified into the relevant categories without any misclassifications among the Normal, Immature, and Mature cataract images. The fact that there are only diagonal values and zero values in off-diagonal places in the matrices proves the absence of false positives and false negatives for all the classes. Thus, the technology of DAG-ResCNN, which has the ability to extract residual features, has been confirmed by these results, and an improvement in decision-making through the hybrid ANN-based classifier of HARAC-Net has also been achieved. The hold-out approach offered perfect classification results, confirming the suggested models for future automated cataract diagnosis.

Table 4: Performance of the proposed model under hold-out method

Model	Class	Evaluation Metrics (%)				
		Specificity	Recall	Precision	F1-Score	Accuracy
DAG-ResCNN	Normal	100	100	100	100	100
	Immature	100	100	100	100	100
	Mature	100	100	100	100	100
	Average	100	100	100	100	100
DAG-ResCNN-ANN (HARAC-Net)	Normal	100	100	100	100	100
	Immature	100	100	100	100	100
	Mature	100	100	100	100	100
	Average	100	100	100	100	100

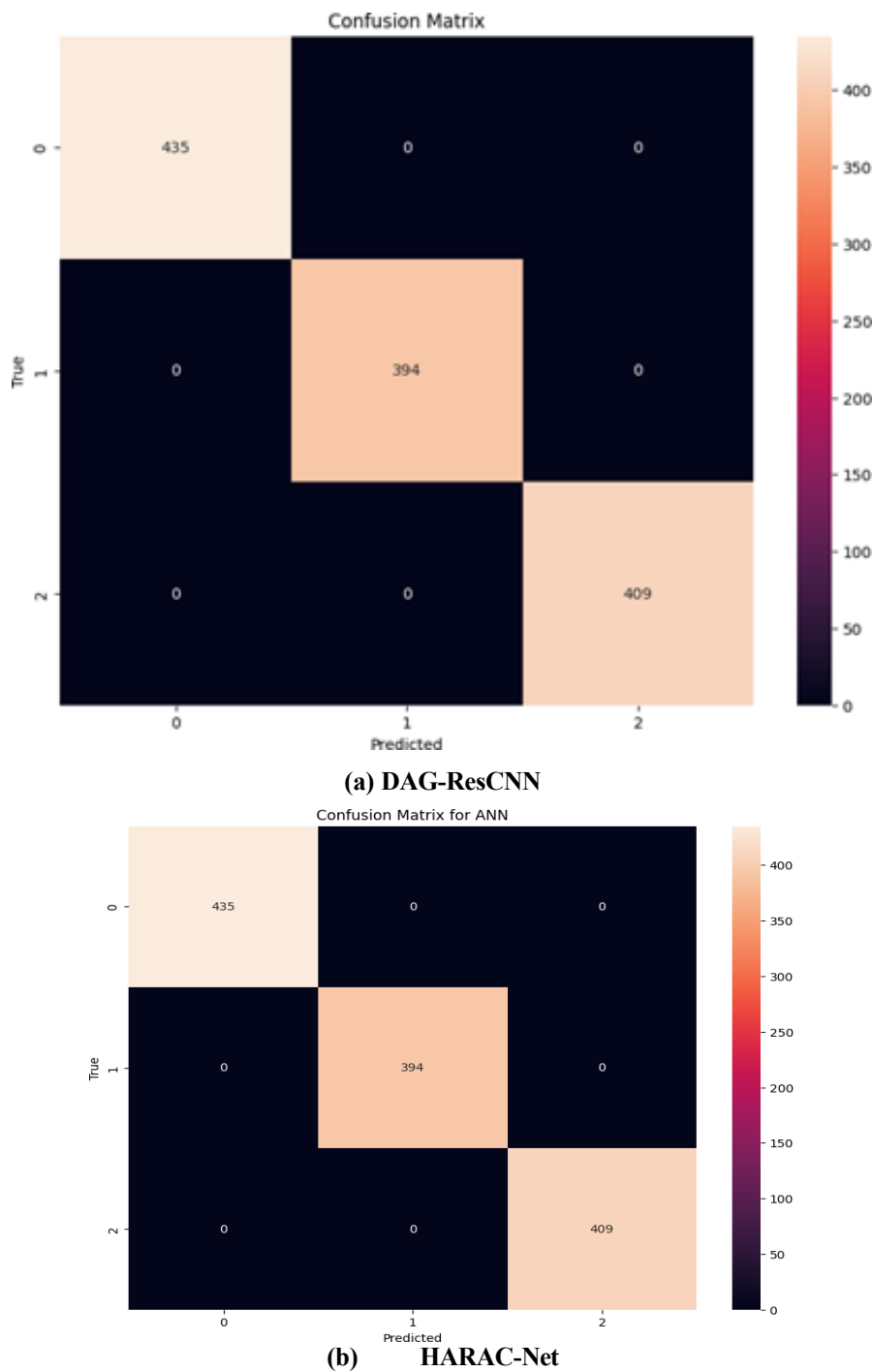


Figure 7: Confusion matrices of the proposed models under hold-out validation (Class 0: Normal, Class 1: Immature, Class 2: Mature)

B. Learning Behavior

In addition to the quantitative metrics presented in Table 4 and the confusion matrices depicted in Figure 7, the learning behavior of the introduced DAG-ResCNN provides additional insight into the training and validation accuracy and loss curves, as shown in Figure 8.

The accuracy curves (Figure 8(a)) reveal exceptionally fast and stable model convergence during the training period. At the start, there was a marked increase in the training and validation accuracies, indicating successful feature

learning from the cataract images. With every epoch, the accuracies continue to increase until they reach one, and the difference between the two remains small throughout the training. The small difference between the accuracy curves of the training and validation sets indicates that the proposed model is good at predicting new data, and there are no signs of overfitting. This type of behavior indicates the very good performance of the attention-guided residual architecture in the area of discriminating features for Normal, Immature, and Mature cataract classification.

Similarly, the loss curves (Figure 8(b)) support the claim that the model optimization is both stable and efficient. There was a significant decline in the training and validation losses during the first few epochs, which signified a rapid error minimization. As the training continued, both curves slowly came together at values

close to zero and remained thus, indicating that the model had reached its optimum. The closeness of the training and validation loss curves guarantees that the proposed framework is robust and can deliver consistent performance, even on unseen samples.

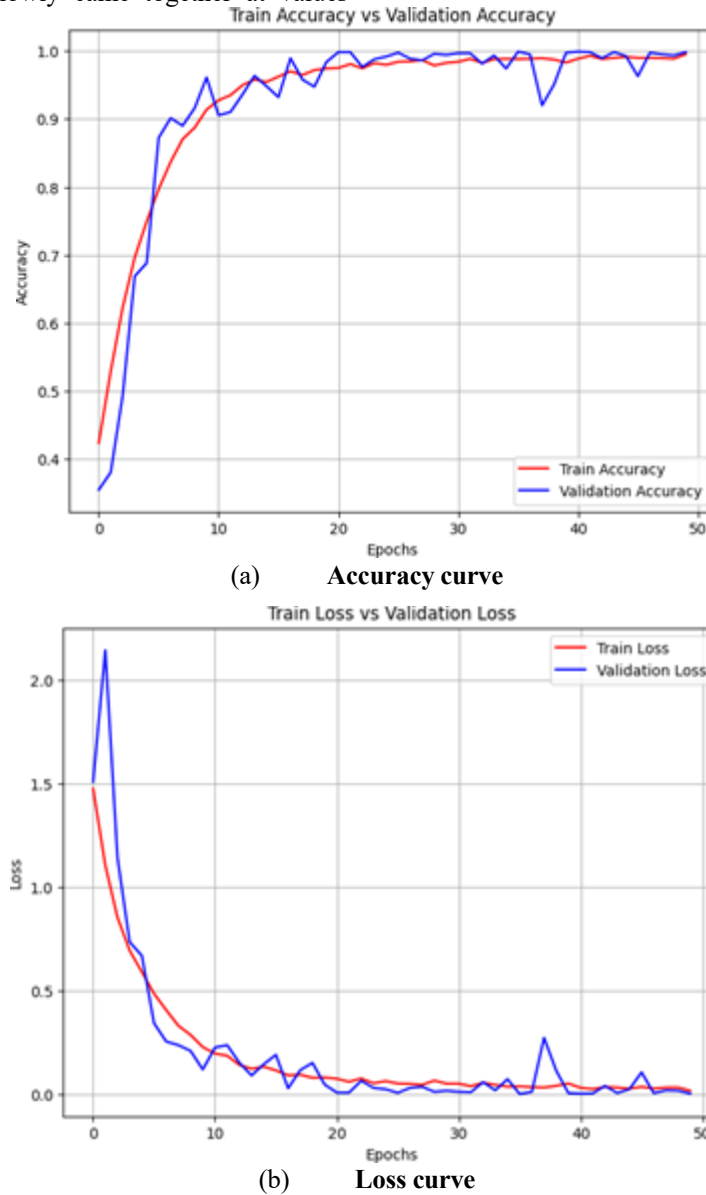


Figure 8: The suggested DAG-ResCNN network accuracy and loss curves

C. Feature Representation

In addition, the proposed model's ability to represent features was examined via feature visualization using principal component analysis (PCA), as shown in Figure 9. The PCA visualization shows the distribution of the deep features extracted by the deep attention-guided residual CNN (DAG-ResCNN) for the different cataract classes: normal (class 0), immature (class 1), and mature (class 2).

The features are shown on the first two principal components (PC1 and PC2) to facilitate the observation of the separation of the classes. The feature clusters of the

different classes are very well separated, and there is little overlap, suggesting that the proposed model is very good at producing not only discriminative but also class-specific representations. A clear and tightly woven cluster of normal cataract images was distinctly separated from the Immature and Mature classes. Similarly, the feature space of the Immature and Mature cataract samples was such that the model's power to detect even the slightest differences in the different stages of cataracts was demonstrated. The clear separation between classes and the very small variances within classes reflect the power of the attention mechanism and residual connections for

feature discrimination.

The PCA-based feature visualization supports the quantitative results and confusion matrix analysis based on its ability to provide visual proof of the strong discriminative power of the learned deep features. This indicates that the proposed DAG-ResCNN and HARAC-Net frameworks not only achieve perfect classification performance when using the hold-out method but also produce feature representations that are robust, easy to interpret, and thus suitable for reliable automated cataract

diagnosis.

4.3. Cross-Validation Analysis of the Proposed Model

The hold-out validation strategy enables an initial assessment of model performance; however, it may not fully capture the generalization ability because of its dependency on a single data split. Consequently, 5-fold cross-validation (5-FCV) was used to comprehensively evaluate the consistency and stability of the proposed DAG-ResCNN and HARAC-Net models by minimizing the bias and variance in the performance evaluation.

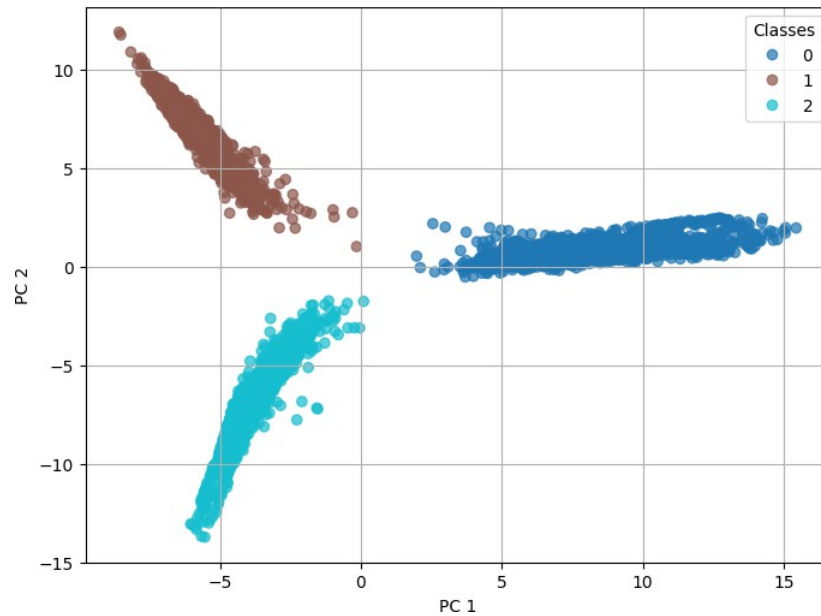


Figure 9: PCA-based feature visualization of cataract classes

A. Quantitative Analysis

The simulation results for the proposed DAG-ResCNN under 5-FCV are listed in Table 5. The simulation results are supported by the confusion matrices shown in Figure 10, which show that a high classification accuracy was achieved for all the folds.

The DAG-ResCNN model performed extremely well in terms of specificity, recall, precision, and F1-score for the five folds, with most of the values tending towards or reaching 100% accuracy, thereby establishing the model's high discriminative power. For Fold 1 (Figure 10(a)), only three samples of the mature class were incorrectly classified as normal, while the rest of the Normal and Immature samples were properly classified. This resulted in a slight decrease in the recall value of the mature class to 99.26%, although the average accuracy remained very high at 99.76%. Likewise, for Fold 2 (Figure 10(b)), a few immature samples (five) were incorrectly classified as normal, as depicted in Figure 10, resulting in a slight decrease in the recall value of the immature class to 98.64%, although the average accuracy was still very high at 99.73%.

It is pertinent to note that Folds 3 and 4 in Figures 10(c) and 10(d), respectively, showed no errors in classification, and their respective confusion matrices captured the three classes without any errors. Therefore, all the

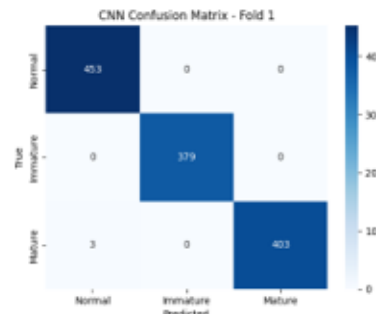
performance metrics—specifically 100% for specificity, recall, precision, F1-score, and accuracy—demonstrate the robustness of the DAG-ResCNN model. In Fold 5 in Figure 10(e), the errors were very limited, with one normal sample classified as mature and four mature samples classified as immature. Despite the limited errors, the model demonstrated a high average accuracy of 99.73%.

In summary, the average performance across the folds yielded an accuracy of 99.86%, which was accompanied by mean $\pm$ standard deviation values of $99.90 \pm 0.09\%$ specificity, 99.78

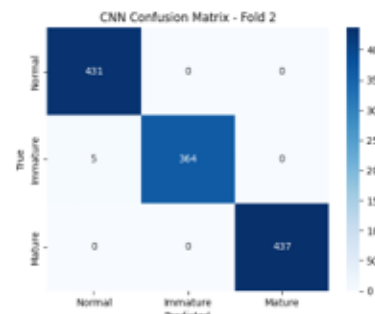
$\pm 0.19\%$ recall, $99.80 \pm 0.18\%$ precision, and $99.80 \pm 0.19\%$ F1-score. The extremely low standard deviation values for all metrics confirm the consistency and generalization ability of the DAG-ResCNN model across different data partitions. The confusion matrices presented in Figure 10 not only visually support these results, but are also characterized by strong diagonal dominance with almost no entries in the off-diagonal zones, indicating that the proposed attention-guided residual architecture efficiently diminishes class overlap. To further improve the classification performance, we extracted features from the dense layer with 50 neurons and employed ANN. The implications are presented in Table 6.

Table 5: Performance evaluation of the DAG-ResCNN under the 5-FCV

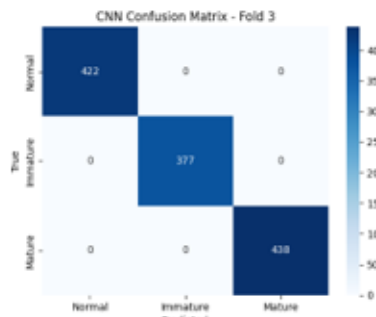
Fold	Class	Evaluation Measures (%)				
		Specificity	Recall	Precision	F1-score	Accuracy
1	Normal	99.65	100	99.34	99.67	99.76
	Immature	100	100	100	100	100
	Mature	100	99.26	100	99.63	99.76
	Average	99.88	99.75	99.78	99.77	99.84
2	Normal	99.56	100	98.86	99.43	99.60
	Immature	100	98.64	100	99.31	99.60
	Mature	100	100	100	100	100
	Average	99.85	99.55	99.62	99.58	99.73
3	Normal	100	100	100	100	100
	Immature	100	100	100	100	100
	Mature	100	100	100	100	100
	Average	100	100	100	100	100
4	Normal	100	100	100	100	100
	Immature	100	100	100	100	100
	Mature	100	100	100	100	100
	Average	100	100	100	100	100
5	Normal	100	99.76	100	99.88	99.92
	Immature	99.53	100	98.97	99.48	99.68
	Mature	99.88	99.10	99.77	99.43	99.60
	Average	99.80	99.62	99.58	99.60	99.73
Mean $\pm$ SD		99.90$\pm$0.09	99.78$\pm$0.19	99.8$\pm$0.18	99.8$\pm$0.19	99.86$\pm$0.13



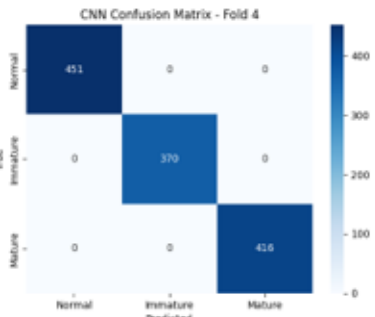
(a)



(b)



(c)



(d)

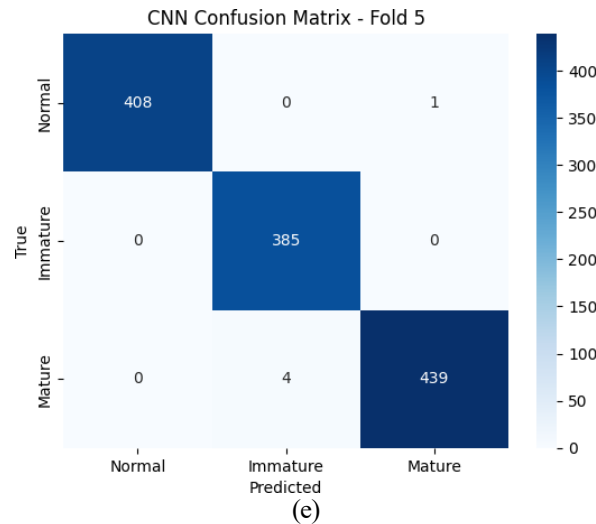


Figure 10: Confusion matrices of the DAG-ResCNN across each fold

The performance of the HARAC-Net model, which merges the feature extractor DAG-ResCNN with an ANN classifier, is listed in Table 6. This performance was evaluated using 5-FCV for the Normal, Immature, and Mature classes. The quantitative results are supported by the confusion matrices shown in Figure 11, which exhibit accurate and stable classification performance for all folds.

HARAC-Net for the five folds performed almost perfectly to perfectly in terms of specificity, recall, precision, F1-score, and accuracy. The confusion matrix in Fold 1 (Figure 11(a)) indicates perfect diagonal dominance, with all samples from the three classes being correctly classified and no misclassifications being present, thus leading to a 100% performance for all evaluation metrics. In Fold 2 (Figure 11(b)), only one mature sample was misclassified as immature, as shown in the confusion matrix. This small error caused a slight reduction in recall for the mature class (99.77%) and a slight decrease in precision for the immature class (99.73%). Nevertheless, the overall average accuracy remained very high at 99.92%, reflecting the strength of the HARAC-Net architecture.

Folds 3 and 4 (Figures 11(c) and 11(d)) also show a

perfect classification performance in that the confusion matrices do not have any off-diagonal elements, while all the metrics achieve 100%. This demonstrates the generalization capability and consistency of HARAC-Net for different datasets. For fold 5 (Figure 11(e)), only a few misclassification instances were observed, such as one instance of a normal cell being classified as mature, two instances of immature cells being classified as mature, and one mature cell being classified as immature. However, the few errors did not affect the recall and precision of the immature and mature classes; however, the model showed an excellent average accuracy of 99.68%.

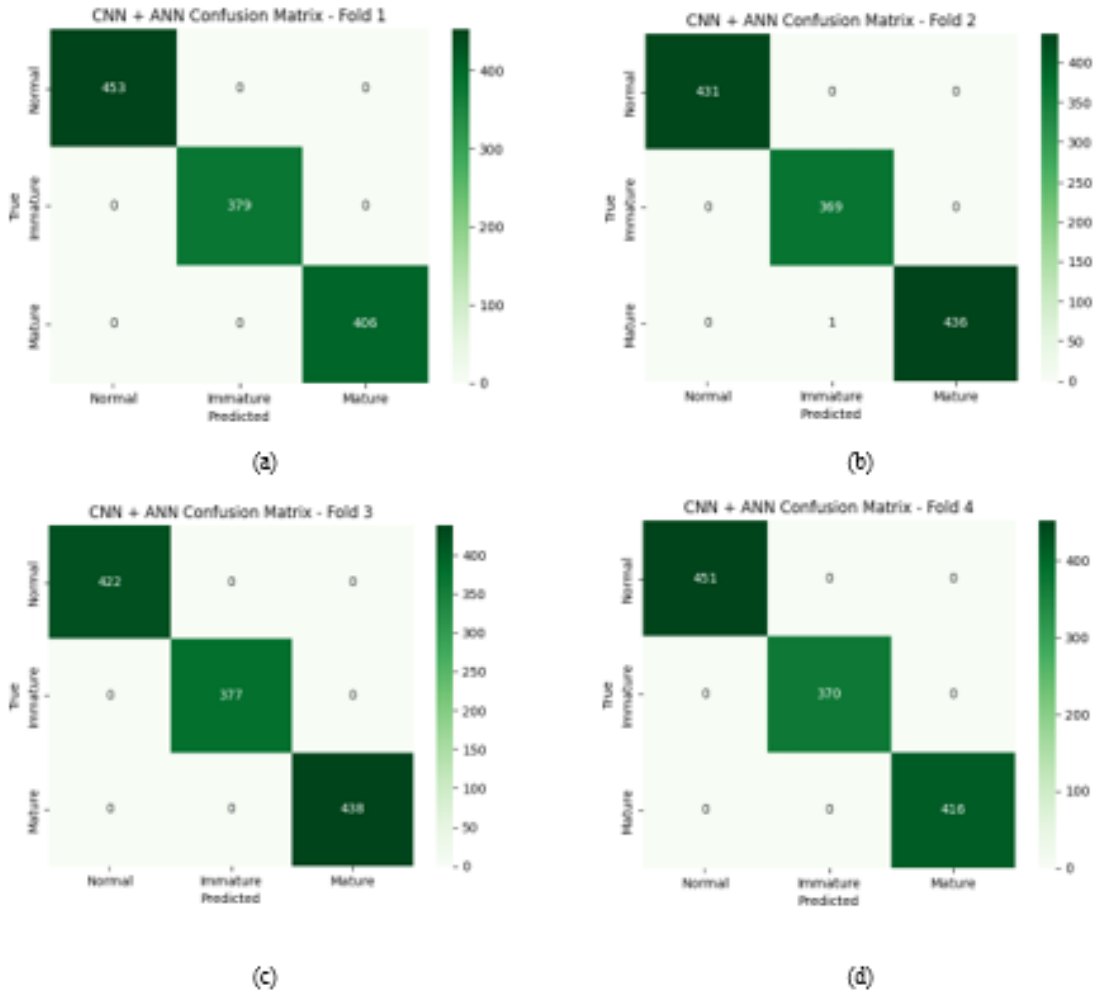
The overall performance of HARAC-Net correlates to a mean accuracy of 99.92%, accompanied by a mean $\pm$ standard deviation of $99.96 \pm 0.07\%$ specificity, $99.92 \pm 0.13\%$ recall,

$99.92 \pm 0.12\%$ precision, and $99.92 \pm 0.11\%$ F1-score. The extremely low standard deviation for all metrics substantiates the stability, reliability, and generalization ability of the proposed hybrid architecture. The confusion matrices in Figure 11 provide visual proof of these results by displaying strong diagonal dominance and almost no misclassification over the folds.

Table 6: Performance evaluation of the HARAC-Net under the 5-FCV

Fold	Class	Evaluation Measures (%)				
		Specificity	Recall	Precision	F1-score	Accuracy
1	Normal	100	100	100	100	100
	Immature	100	100	100	100	100
	Mature	100	100	100	100	100
	Average	100	100	100	100	100
2	Normal	100	100	100	100	100
	Immature	99.88	100	99.73	99.86	99.92
	Mature	100	99.77	100	99.88	99.92
	Average	99.92	99.96	99.91	99.91	99.92
3	Normal	100	100	100	100	100
	Immature	100	100	100	100	100
	Mature	100	100	100	100	100
	Average	100	100	100	100	100

4	Normal	100	100	100	100	100
	Immature	100	100	100	100	100
	Mature	100	100	100	100	100
	Average	100	100	100	100	100
5	Normal	100	99.76	100	99.88	99.92
	Immature	99.88	99.48	99.74	99.61	99.68
	Mature	99.62	99.77	99.33	99.55	99.68
	Average	99.83	99.67	99.69	99.68	99.68
Mean ± SD		99.96 ± 0.07	99.92 ± 0.13	99.92 ± 0.12	99.92 ± 0.11	99.92 ± 0.13



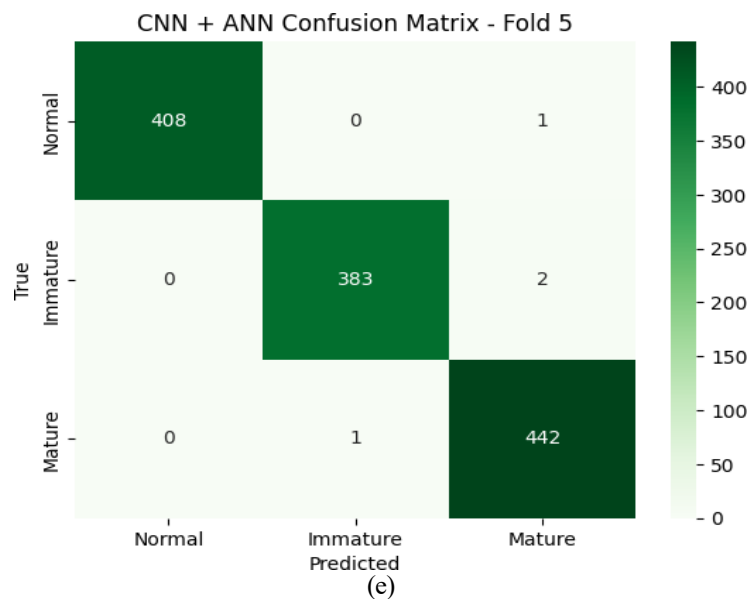


Figure 11: Confusion matrices of the HARAC-Net across each fold

B. Comparative Analysis of DAG-ResCNN and HARAC-Net

Figure 12 shows a thorough comparison of DAG-ResCNN and HARAC-Net regarding all significant evaluation metrics with mean $\pm$ standard deviation (%) over 5-FCV. The error bars indicate the extent of variability for each metric across the different folds, thus supporting statistical transparency and reproducibility of the results.

Figure 12 evidently reveals that HARAC-Net outperforms DAG-ResCNN in terms of all the evaluation metrics and also shows a lower standard deviation, which implies higher stability and generalization. HARAC-Net scored higher mean values across the board for recall, precision, F1-score, and accuracy, thus validating the efficiency of the integration of an ANN-based decision layer on a deep attention-guided residual CNN. The largest gains were particularly evident in the recall and F1-score, which are the most important metrics to consider from a clinical perspective. The increased recall indicates that HARAC-Net is good at catching all true cases (TPs), thereby minimizing false negatives (FNs), whereas the improved F1-score shows that there indicates a good balance between precision and recall. The shrinking error bars also suggest that there was not much variation in the model performance across different folds, thus confirming the robustness of the proposed hybrid architecture.

From a methodological perspective, the very close confidence range near the top performance limit ($>99.9\%$) illustrates that both models performed excellently. Nevertheless, the consistent upward shift of HARAC-Net across all metrics strongly indicates that the ANN layer contributes to the creation of more refined decision boundaries and the enhancement of class separability, particularly in difficult cases with slight inter-class similarities.

Even small enhancements at high accuracy levels are important in clinical and real-world deployment. The increased recall and F1-score, accompanied by a lower variance, imply that HARAC-Net is a more trustworthy and faithful framework for automated screening and decision-support systems for the diagnosis of cataract-related diseases than the other models. Consistency that has improved through folds minimizes the possibility of performance drop when applied to new data, which is very important for clinical acceptance and large-scale application.

Finally, Figure 12 confirms through visual and quantitative means that while DAG-ResCNN provides a solid baseline, HARAC-Net reaches greater performance, robustness, and clinical reliability, thus making it a better candidate for practical and scalable AI-assisted diagnostic solutions.

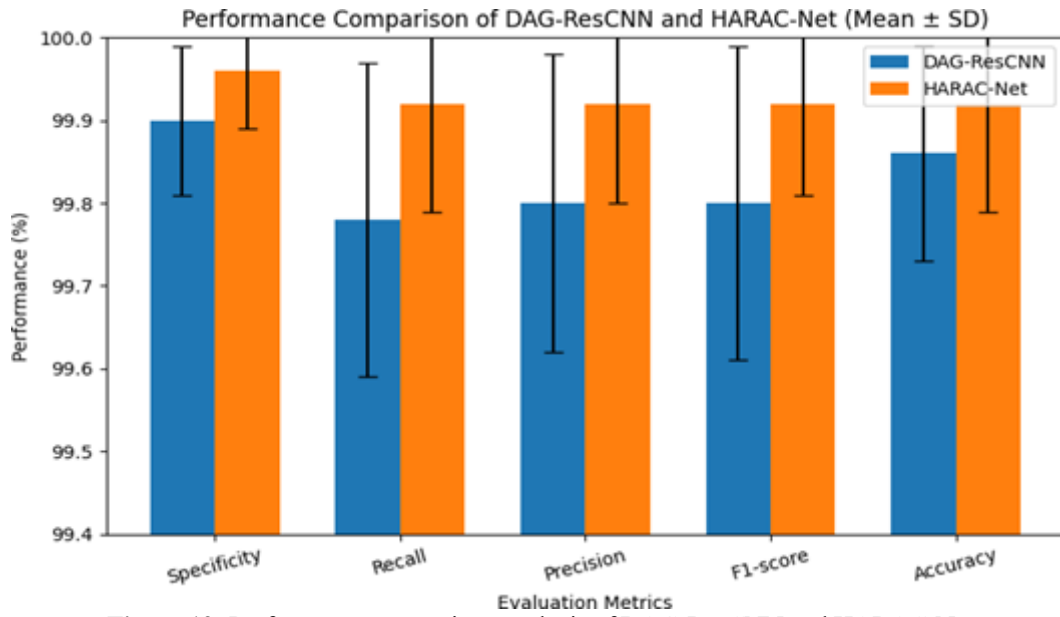


Figure 12: Performance comparison analysis of DAG-ResCNN and HARAC-Net

4.4. Feature Interpretability

Feature interpretation was validated by the alignment between the highlighted Grad-CAM and clinically important areas. The Grad-CAM technique has proven to be an indispensable tool for interpreting deep learning-based cataract assessment systems, not only by producing reliable classification outcomes but also by augmenting clinical trust via explainability. For a better understanding, the key insights of Grad-CAM are listed in Table 8.

Formally, the Grad-CAM saliency map for class c with respect to feature map A^k of the final convolutional layer is computed through three steps. First, the class-discriminative weights α^k for each feature map channel k are obtained by global average pooling of the gradient of the class score S^c with respect to A^k , as expressed in Eq. (23):

$$\alpha_c^k = \frac{1}{Z} \sum_i \sum_j \frac{\partial S^c}{\partial A_{ij}^k} \quad (23)$$

where Z is the total number of spatial positions in the feature map. Second, the weighted combination of the activation maps is computed, and a ReLU is applied to

retain only the positive influences on the class prediction, as expressed in Eq. (24):

$$L_{\{Grad-CAM\}}^c = ReLU(\sum_k \alpha_c^k \cdot A^k) \quad (24)$$

The resulting heatmap $L^c_{\{Grad-CAM\}}$ is upsampled to the original image resolution and overlaid on the input ocular image to highlight cataract-relevant regions (lens opacity, texture degradation). The ReLU operation discards spatial

locations that have a negative effect on the predicted class, ensuring that only class-specific discriminative regions are visualized.

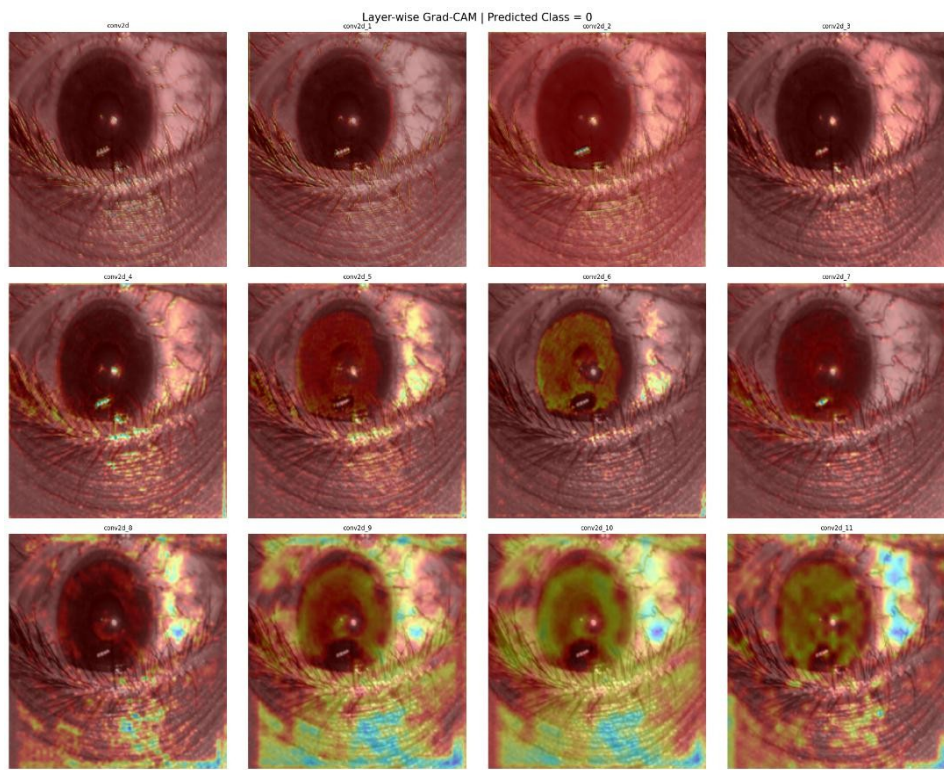


Figure 13: Interpretation of Normal eye image using Grad-CAM across layer-wise

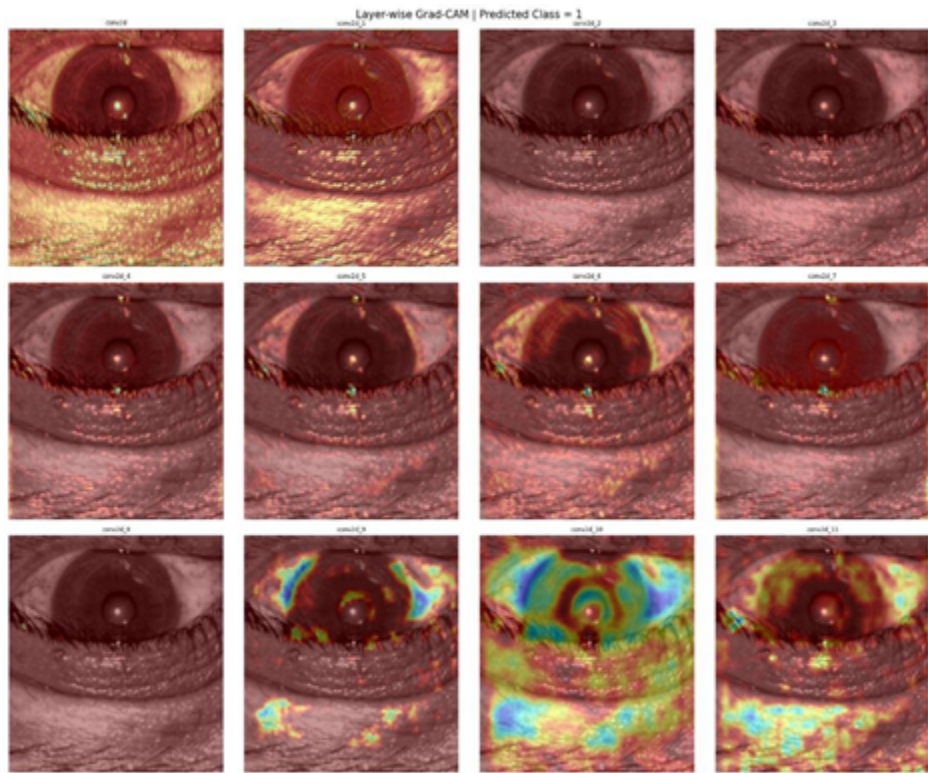


Figure 14: Interpretation of immature cataract eye image using Grad-CAM across layer-wise

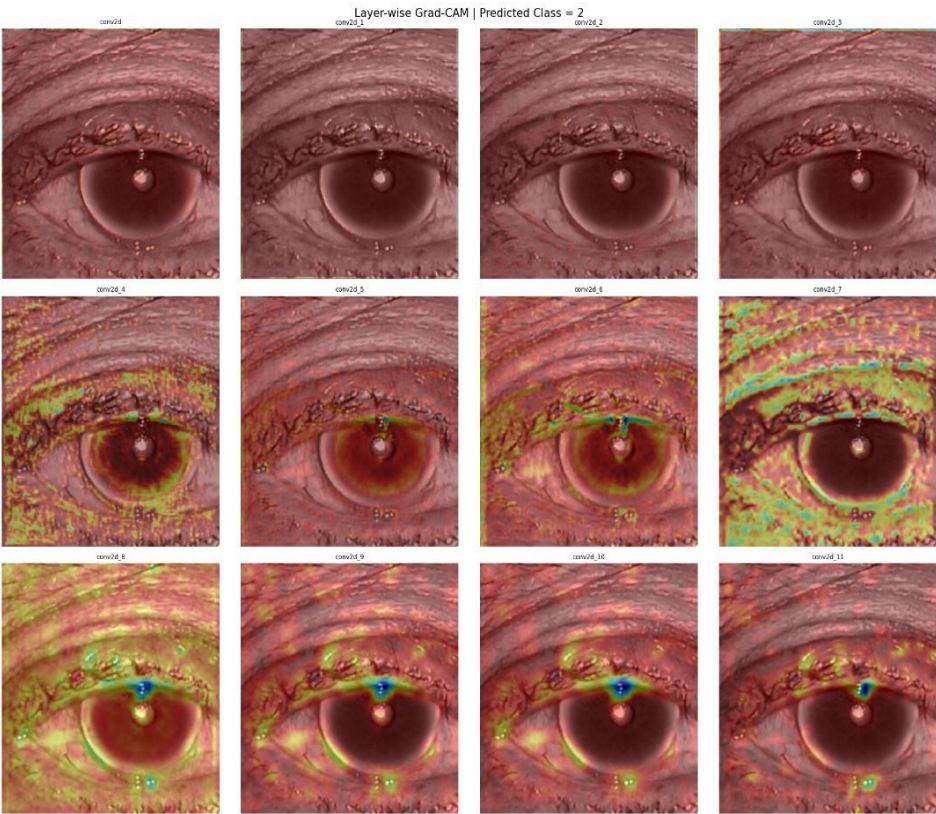


Figure 15: Interpretation of mature cataract eye image using Grad-CAM across layer-wise

Table 8: Key insights of layer-wise Grad-CAM across various classes of eye cataract images

Aspect	Normal	Immature	Mature
Early Layers	Generic edges	Generic edges	Early lens cues
Mid Layers	Weak lens focus	Localized lens focus	Strong lens focus
Deep Layers	Diffuse attention	Moderate central activation	Dense, dominant activation
Clinical Meaning	Transparent lens	Partial opacity	Severe opacity

4.5 Comparative Analysis

To objectively evaluate the efficiency of the proposed HARAC-Net, a comparative analysis was conducted between the proposed model and other state-of-the-art models for cataract detection and classification presented in the literature. Because many approaches were initially developed for two-class classifications or had differences in datasets and setups used, these frameworks had to be redesigned and implemented to conduct a comparative analysis of the three-class cataract classification framework under the same setup as the proposed approach.

The results of the comparative analysis of the proposed model with other cataract detection and classification models are listed in Table 9. As demonstrated by Agustin et al. [13] and Irshad Khan et al. [18], some models based on deep CNN architectures have achieved outstanding

results. Nevertheless, there are still many models that have high complexity, many trainable parameters, and are tested only in two-class classification problems. Furthermore, the research conducted by Al Sariera et al. [16] showed that machine learning models have poor generalization ability.

Conversely, the proposed HARAC-Net, which combines residual learning with dual attention mechanisms (SE and CBAM), outperformed all metrics of specificity, precision, recall (sensitivity), F1-score, and overall accuracy. The proposed model had an accuracy of $99.92 \pm 0.13\%$ (mean $\pm$ SD) using 5-fold cross-validation (5-FCV) and 100% accuracy using the hold-out method, with balanced precision and sensitivity. This shows good discrimination between normal, immature, and mature cataract stages. In addition, the model performed well without cross-validation, with perfect scores for classification, thus demonstrating the stability of the proposed architecture.

Table 9: Performance comparison of proposed and existing deep learning models for three-class cataract classification

Method / Model	Dataset Split	No. of Trainable Parameters	Specificity (%)	Recall (%)	Precision (%)	F1-Score (%)	Accuracy (%)
Customized CNN model [13]	80/20	22.25 M	99.22	98.39	98.52	98.44	98.46

Hybrid CNN [15]	80/20	7.9 M	95.43	95.12	95.12	95.23	96.25
ML Classifiers [16]	80/20	Depends on Classifier	95.41	90.87	90.87	90.87	90.87
Resnet50 [18]	80/20	23.5 M	100	100	100	100	100
Customized CNN model [20]	80/20	19.56 M	99.88	99.76	99.75	99.75	99.76
Customized CNN model [29]	80/20	NP	NP	NP	NP	NP	98.3
MobileNetV2 [30]	80/20	3.5 M	NP	NP	NP	NP	94
CNN with LeNet [31]	80/20	NP	NP	NP	NP	NP	93.33
HARAC-Net	5-FCV	2 M	99.96 ± 0.07	99.92 ± 0.13	99.92 ± 0.12	99.92 ± 0.11	99.92 ± 0.13
	80/20		100	100	100	100	100

*NP: Not Provided and * The format 99.96 ± 0.07 shows mean $\pm$ SD

Compared with other customized CNN architectures [15, 20, 29] and light CNN models such as MobileNetV2 [30] and LeNet-based CNN architectures [31], the proposed system has a good trade-off between model complexity and classification accuracy. The addition of attention-driven feature learning helps the network focus on important regions of opacity and texture, thus improving its discriminative ability while maintaining interpretability.

In general, the comparative results show that the proposed HARAC-Net model outperforms the current state-of-the-art methods on the three-class cataract dataset, indicating that the proposed model is suitable for large-scale cataract screening and computer-aided ophthalmic diagnosis systems.

5. CONCLUSION

In this study, a hybrid attention-aware residual deep learning model, named HARAC-Net, was developed for the automatic classification of cataract stages from ocular images. The proposed technique combines a residual CNN model with a squeeze-and-excitation (SE) and Convolutional Block Attention Module (CBAM) to focus on important cataract features corresponding to opacity and texture. Additionally, the proposed technique combines deep attention-driven features with an artificial neural network (ANN) classifier. A series of experiments was performed on a three-class cataract image dataset, including normal, immature, and mature cataracts. Performance analysis on this dataset shows that HARAC-Net achieves better results on all performance metrics—Accuracy, Precision, Recall, Sensitivity, Specificity, and F1 measure—compared to the redesigned existing state-of-the-art models because of the improved generalization ability of HARAC-Net.

In addition to the quantitative improvements, the embedded Grad-CAM visualization tool provides useful explanations by highlighting the areas in the ocular images that impact the decision-making process of the model. This is particularly important in the medical environment, as it improves trust and enables a

collaborative approach between the diagnostic system and ophthalmologists. In conclusion, the experimental results show that HARAC-Net outperforms the state-of-the-art methods based on a single evaluation metric for a three-class cataract image dataset, providing a favorable trade-off between accuracy, interpretability and efficiency. These properties are significant for the practicality of the proposed framework in large-scale cataract screening and computer-assisted ophthalmic diagnosis.

Future work will include generalizing the proposed framework to multi-center and large-scale datasets and will aim to develop a light version of the model that can be applied in real-time, focusing on the longitudinal analysis of disease progression.

REFERENCES

1. World Health Organization (WHO). "World report on vision." (2019).
2. Flaxman, Seth R., et al. "Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis." *The Lancet Global Health* 5.12 (2017): e1221-e1234.
3. Steinmetz, Jaimie D., et al. "Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study." *The Lancet Global Health* 9.2 (2021): e144-e160.
4. Chylack, Leo T., et al. "The lens opacities classification system III." *Archives of ophthalmology* 111.6 (1993): 831-836.
5. de Souza, Rodrigo Guimaraes, et al. "Association of optical cataract indices with cataract severity and visual function." *International ophthalmology* 42.1 (2022): 27-33.
6. Grewal, Dilraj S., Gagandeep S. Brar, and Satinder Pal Singh Grewal. "Correlation of nuclear cataract lens density using Scheimpflug images with Lens

- Opacities Classification System III and visual function." *Ophthalmology* 116.8 (2009): 1436-1443.
7. Lim, Dong Hui, et al. "Measurement of lens density using Scheimpflug imaging system as a screening test in the field of health examination for age-related cataract." *British Journal of Ophthalmology* 99.2 (2015): 184-191.
8. Jabber, Abdullah Ali Jabber, and Hayder Mahdawi. "A Comprehensive Review of Deep Learning Techniques for Detecting and Classifying Cataracts and Glaucoma." *Al-Furat Journal of Innovations in Electronics and Computer Engineering* 4.2 (2025): 1-11.
9. Imran, Azhar, et al. "Automated identification of cataract severity using retinal fundus images." *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 8.6 (2020): 691-698.
10. Zheng, Ce, et al. "Development and validation of deep learning algorithms for automated eye laterality detection with anterior segment photography." *Scientific Reports* 11.1 (2021): 586.
11. Dong, Yanyan, et al. "Classification of cataract fundus image based on deep learning." 2017 IEEE international conference on imaging systems and techniques (IST). IEEE, 2017.
12. Ting, Daniel Shu Wei, et al. "Artificial intelligence and deep learning in ophthalmology." *British Journal of Ophthalmology* 103.2 (2019): 167-175.
13. Agustin, Sarah, Eka Novelia Putri, and Ichwan Nul Ichsan. "Design of A Cataract Detection System based on The Convolutional Neural Network." *Jurnal ELTIKOM: Jurnal Teknik Elektro, Teknologi Informasi dan Komputer* 8.1 (2024): 1-8.
14. PL, Lahari, et al. "CSDNet: a novel deep learning framework for improved cataract state detection." *Diagnostics* 14.10 (2024): 983.
15. Nguyen, Van-Viet, and Chun-Ling Lin. "Enhancing Cataract Detection through Hybrid CNN Approach and Image Quadrature: A Solution for Precise Diagnosis and Improved Patient Care." *Electronics* 13.12 (2024): 2344.
16. Al Sariera, Jamil, Esra'A. Mahmoud, and Thamer Mitib Al Sariera. "Automated Cataract Detection and Classification Using Random Forest Classifier in Fundus Images." *International Journal of Intelligent Engineering & Systems* 17.4 (2024).
17. Mahmood, Salwa Shakir, Sihem Chaabouni, and Ahmed Fakhfakh. "Improving Automated Detection of Cataract Disease through Transfer Learning using ResNet50." *Engineering, Technology & Applied Science Research* 14.5 (2024): 17541-17547.
18. Khan, Irshad, et al. "Enhancing ocular health precision: Cataract detection using fundus images and ResNet-50." *ICCK Transactions on Intelligent Systematics* 1.3 (2024): 145-160.
19. Santoso, Ilham, Ayub Michaelangelo Manurung, and Egia Rosi Subhiyakto. "Comparison of ResNet-50, EfficientNet-B1, and VGG-16 Algorithms for Cataract Eye Image Classification." *Journal of Applied Informatics and Computing* 9.2 (2025): 284-294.
20. Ramdhani, Fitra Rizki, et al. "Feature Extraction in Eye Images Using Convolutional Neural Network to Determine Cataract Disease." *International Journal of Engineering and Computer Science Applications (IJECSA)* 4.2 (2025): 81-90.
21. Ramakrishnan, Akshay Bhuvaneswari, et al. "A Hybrid Deep Learning Paradigm for Robust Feature Extraction and Classification for Cataracts." *Applied AI Letters* 6.2 (2025): e113.
22. Syahda, Tabina Shafa Nabila, Mochamad Nizar Palefi Ma'ady, and Berlian Rahmy Lidiawaty. "A Comparative Study of VGG16, MobileNet, and Ensemble Methods on Fundus Image-Based Cataract Detection." 2025 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS). IEEE, 2025.
23. Islam, Md Sariful, et al. "Cat-CNN: Human Eye Cataract Detection from Color Fundus Photograph with Deep CNN with Optimized Cascaded Network." 2024 IEEE International Conference on Signal Processing, Information, Communication and Systems (SPICSCON). IEEE, 2024.
24. Fahmi, Fahmi, and Hizkia Estiko. "Cataract detection using deep learning in teleophthalmology systems." 2024 14th International Conference on Electrical Engineering (ICEENG). IEEE, 2024.
25. Kaushik, Pratham, and Pooja Sharma. "Cataract Detection Using Transfer Learning on Fundus Images: A Deep Learning Approach for Automated Ophthalmic Diagnosis." 2025 International Conference on Automation and Computation (AUTOCOM). IEEE, 2025.
26. <https://www.kaggle.com/datasets/mohammedgamal37130/eye-cataractmature-immature-normal>
27. Reddy, K. Rasool, et al. "Design of a medical decision-supporting system for the identification of brain tumors using entropy-based thresholding and non-local texture features." *Frontiers in Human Neuroscience* 17 (2023): 1157155.
28. Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.
29. Lakshmi, K. Mohana, et al. "Leveraging

convolutional neural network for cataract classification in retinal fundus images." *Advancing Innovation through AI and Machine Learning Algorithms*. CRC Press, 2025. 553-558.

30. Nair, Shruthi Viju, and Prasanna Shete. "Mobile Application for Cataract Detection Using Convolution Neural Network." *2023 International Conference on Network, Multimedia, and Information Technology (NMITCON)*. IEEE, 2023.
31. Labib, Radimas Putra Muhammad Davi, et al. "Convolutional neural network for cataract maturity classification based LeNet." *Jurnal Varian* 5.2 (2022): 97-106.