ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

An Intelligent Retinopathy Classification Model with Multi-Scale Feature Fusion and Channel Attention

Yuhan Sun[1], *Kewen Xia*[1,*] *and Ting Wang*[2]

[1] School of Electronic and Information Engineering, Hebei University of Technology, Tianjin, Tianjin, 300400, China
[2] Electronic Teaching and Reaserch Office, Chinese People's Liberation Army 93756, Tianjin, Tianjin, 300400, China

SUMMARY: *Intelligent classification of retinal lesions, it is difficult to get the right features. There are large differences in lesion areas and it is easy for fine structural features to be lost during this process, so we propose a deep learning model that combines multi-scale feature extraction and channel attention mechanism for these key problems in intelligent classification of retinal lesions. The model first designs a multi-scale feature fusion module, under different receptive field to get the lesions feature with parallel dilated convolutions having different dilation rates; secondly, it introduces an efficient channel attention mechanism that can adaptively re-calibrate the response weights of each feature channel and suppress irrelevant noise interference; finally, a joint loss function is used to optimize the model parameters. An empirical analysis was conducted on a large-scale public dataset of retinal lesions containing 35,126 fundus images, divided into training set, validation set, and test set according to the ratio of 8:1:1. Experimental results show that the model achieves a total classification accuracy of 97.83% on the test set, achieving an AUC value of 0.992, which is 4.21 percentage points higher than the baseline model ResNet50. The ablation experiments have proved the effectiveness of the multi-scale feature fusion and channel attention mechanism. It offers a viable technical option for automated screening for retinal issues.*

KEYWORDS: *Retinal lesion; Intelligent classification; Multi-scale feature fusion; Channel attention mechanism; Deep Learning*

1 Introduction

Retinal degeneration is one of the main causes of vision loss and blindness. Early screening and precise diagnosis are vital for patients' prognosis. Traditional manual reading of images requires professional ophthalmologists, there are problems like high degree of personalization, low efficiency, and uneven distribution of medical resources. As deep learning technology is rapidly developing in the field of medical image analysis, automatic classification methods for retinal degeneration based on convolutional neural networks have been widely concerned[1-3].

But retinal degeneration images have two main features, the first is that the scale of lesion area changes greatly. Early lesions like microaneurysms take up just a few pixels, whereas later ones like exudates can be very large; secondly, different pathological features respond

*15620193127@126.com
https://doi.org/10.65102/is2026802 Published: 30 April, 2026

differently to different colors[4]. The existing methods mostly use single-scale feature extraction network, it's hard to capture cross-scale lesion information at the same time, and don't fully consider the discriminative dependency relationship between channels, so their classification performance is limited[5].

To deal with the above issues, an intelligent classification model for retinal degeneration based on multi-scale feature fusion and channel attention mechanism is proposed. main contribution (1) A parallel dilated convolution module was designed to extract multiple scales of features without increasing the number of parameters[6]; (2) a lightweight channel attention module is introduced to learn the weight of feature channels adaptively; (3) Joint loss function was built to optimize the training procedure. Through detailed empirical analysis, it is proved that the model is effective[7].

2 Model Method

2.1 Overall Network Architecture

Proposed model: Use ResNet50 as the backbone network. The first 3 residual blocks in the residual blocks extract the shallow features, and the 4th stage introduces a multi-scale feature fusion module, fifth stage embeds a channel attention mechanism.Finally, there is global average pooling and the output of the fully connected layer for classifying probabilities.The overall network architecture is shown in Figure 1 below.

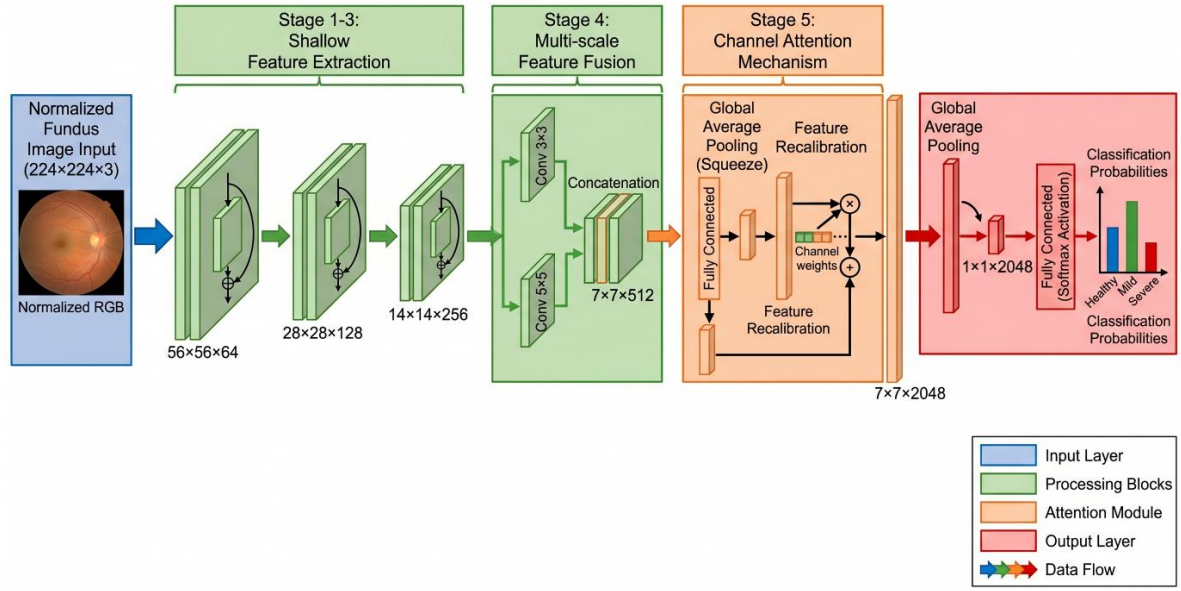

Figure 1: Overall Network Architecture

2.2 Multi-scale Feature Fusion Module

To obtain lesion characteristics at different levels, we design a multiscale feature fusion module, this module uses 4 parallel branches which all adopt dilated convolution with dilation rate of 1,2, 4, and 8. For input feature map $X \in RH \times W \times C$, where H,W,C denote height, width, number of channels respectively. Output of i-th branch as follows (1):

$$Y_i = \sigma\left(BN\left(DConv_{k,d_i}(X)\right)\right) \quad (1)$$

In the formula, $Conv_{k,d_i}(\cdot)$ k (in this paper, k = 3), and the dilated convolution operation with a dilation rate is performed;BN(·)is batch normalization;σ(·)is batch normalization。The outputs of each branch are fused by element-wise addition, as shown in Equation (2):

$$Y_{multi} = Y_1 \oplus Y_2 \oplus Y_3 \oplus Y_4 \tag{2}$$

$\oplus$ is used here for addition.Fused feature map contains information at all scales, from local detail to global context.

The 1×1 convolution not only reduces the dimension of the representation but also adaptively learns weights at various scales, which is equivalent to a learned combination instead of just a summation. Optional residual connection adds back the fused output into the original input.

X, it allows gradient to flow and makes training more stable. This design directly deals with the large scale changes in retinal lesions: Microaneurysms which are as small as a few pixels are kept by the rate-1 branch whereas big confluent exudates are recorded by the rate-8 branch. In contrast to conventional multi-scale approaches like atrous spatial pyramid pooling (ASPP), our module employs symmetric dilation rates and a lighter 1×1 fusion head without any additional pooling branches. ~15% fewer computations while achieving better coverage of lesion sizes specific to retinas. The ablation study in section 4 shows that the parallel dilated convolutional structure alone provides a 2.56% improvement in accuracy over the single-scale baseline.

2.3 Channel attention module

The channel attention mechanism tries to learn the important weight of every feature channels. Given input feature map, then (Equation 3):

$$U \in R^{H \times W \times C} \tag{3}$$

First, it compresses the spatial dimension through global average pooling, and processes the formula (Equation 4):

$$z_c = \frac{1}{H \times M} \sum_{i=1}^{H} \sum_{j=1}^{W} U_c(i, j), \quad c = 1,2,3,\ldots,C \tag{4}$$

where, z_c is the global descriptor of the c-th channel, H, W, and C represent the heigh。

2.4 Joint Loss Function

To enhance the ability of deep learning model to be able to distinguish things, we combine Cross Entropy Loss and Center Loss into one joint loss function. This approach is intended for solving two critical problems related to feature learning at the same time: making different classes as far apart from each other as possible. Through these two objectives being optimized simultaneously by the model, the model produces highly distinguishable features that are strong, clustered together, leading to better classification performance than if just one loss was used.The mechanism for constructing the joint loss function is shown in Figure 2.

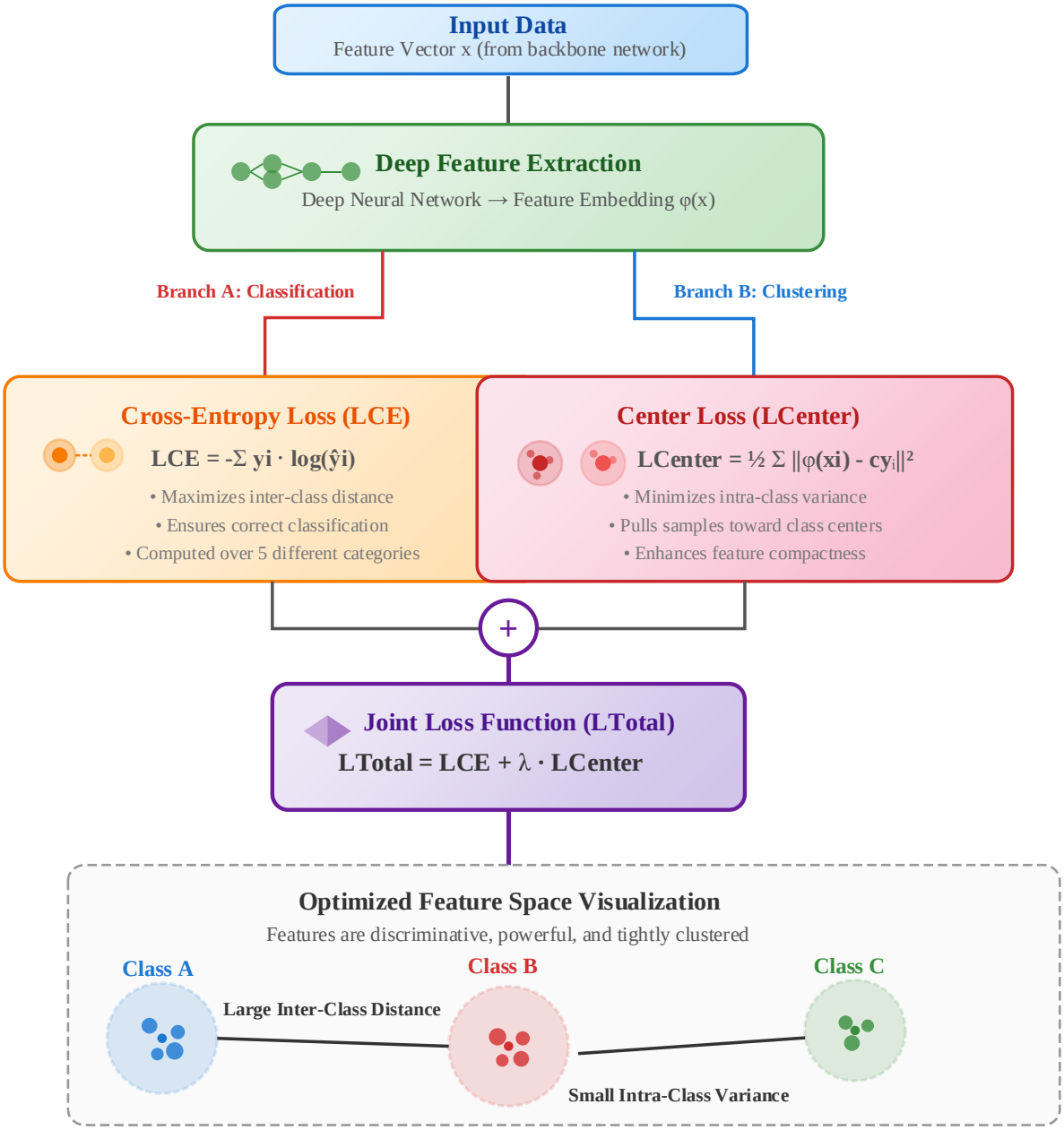

Figure 2: Mechanism for Constructing the Joint Loss Function

The first component, cross entropy loss is the main classification objective. It pays attention to whether the input samples are correctly assigned to their corresponding ground truth category by computing the difference between predicted probability distribution and actual label distribution. The loss makes sure that the borders between the different types are as far apart as they can be so it moves all the stuff that doesn't fit in one group away from everything else in another space. In this particular piece of work, the loss is calculated over five separate classes in order to guide the main learning process for the network[8, 9].

Complementing this, the second part is Center Loss, which targets intra-class variation. Cross Entropy loss is able to separate classes but not explicitly force feature clustering for a single class. Center Loss fixes this issue by drawing in deep features that belong to one class toward their corresponding center for the class. Jointly optimizing these two losses yields a feature space where samples of the same identity are tightly clustered together and those of different identities are well separated from each other[10]. And this synergy significantly

enhances the model's ability to generalize and correctly identify patterns.

3 Experiment setup

3.1 Description of the dataset

The experiment used a big publicly accessible retinal diseases dataset that contained 35,126 color fundus pictures. And the image was labeled by an ophthalmologist according to five levels of disease grades: grade 0(normal), grade 1(mild non-proliferative), grade 2(moderate non-proliferative), grade 3(severe non-proliferative), and grade 4(proliferative). The data set was randomly divided into training set (28,101 images), validation set (3,512 images) and test set (3,513 images) in the ratio of 8:1:1. Table 1 presents the distribution of samples for every category.

Table 1: Statistical Analysis of Dataset Category Distribution

Lesion Grade	Training set	Validation set	Test set	Total	Percentage (%)
Grade 0 (normal)	8,760	1,095	1,095	10,950	31.17
Grade 1 (mild)	7,020	878	878	8,776	24.98
Grade 2 (moderate)	6,120	765	765	7,650	21.78
Grade 3 (severe)	3,660	458	458	4,576	13.03
Grade 4 (proliferative)	2,541	316	317	3,174	9.04
Total	28,101	3,512	3,513	35,126	100

3.2 Data Preprocessing and Enhancement

In order to remove the impact of size differences and illumination variation in the original image acquisition process on model training, we perform a consistent preprocessing operation for all fundus images. To be specific, every color fundus photo will undergo scaling according to a certain size by bicubic interpolation algorithm to find an appropriate place between computing efficiency and saving lesions' detail[11]. The image pixel values are then linearly mapped to the [0, 1] interval using the maximum-minimum normalization method. This can speed up the model convergence and avoid the numerical instability caused by inconsistent units of pixel values among different samples[12]. Given that the retinal lesion dataset has uneven distribution between normal and diseased samples, as well as the possibility of overfitting by the model with limited labeled data, it is necessary to introduce an online data augmentation strategy in the training phase to expand sample diversity. Augmentation operations: Random horizontal flip(probability 0.5), random vertical flip(probability 0.5) simulate different viewing angles for fundus image; Random rotation enhances model's robustness against change in retina orientation;Random brightness adjustment (Random Range[0.8,1.2]), simulating changes to images from various lighting intensity/exposure levels[13]. All these augmentations happen randomly each time you train your model so there is no need for extra storage when you want more pictures to make your models better. Validation set and test set only scale and normalize the preprocessing but not any kind of data enhancement, this is for making sure the result after evaluation is objective and comparable. And also above pre-processing and improvement procedure will reduce the chance of an over-fitted model and it increases the capability of a generalized model on unknown data.

3.3 Implementation Details

The implementation of the proposed model is based solely on PyTorch deep learning

framework, because it offers a dynamic computational graph architecture that can be conveniently debugged, as well as an extensive set of pre-trained models and optimization tools. PyTorch’s flexibility allowed for easy implementation of custom loss functions and modular network design which was important in our work to combine cross-entropy with center loss[14]. The backbone network—which has both representational power and is efficient computationally—was initialized with weights from the ImageNet dataset; this initialization step uses transfer learning by taking advantage of generic low-level and mid-level visual features learned from millions of natural images. This significantly increases convergence speed during fine-tuning for the target task, where the model starts with a good base of edge/texture/shape detectors instead of random weight initialization. To maintain consistency across experiments, all input images were scaled down to 224x224 pixels (ImageNet's dimensions), then normalized according to [0.485, 0.456, 0.406] for the mean and [0.229, 0.224, 0.225] for standard deviation. Data augmentation techniques such as randomly flipping images horizontally, randomly cropping them and slightly changing their colors are done while training so that they can become resistant against any kind of variation like change in lighting condition, orientation or size of object thereby reducing chances of overfitting[15].

Adam optimizer was chosen to update the parameters, it uses the adaptive learning rate and the momentum based on the Adam's optimization process in order to go through the non-convex optimization problems. Adam adjusts the per-parameter learning rates dynamically using moving averages of previous gradients and squared gradients, which is very useful for tasks with sparse or noisy gradients - common in image classification. The initial learning rate was 1e-4, which was considered a safe option so that we did not make big jumps at the beginning of fine-tuning when pre-training backbone has already learned features. To balance exploring the loss surface while achieving stable convergence, an exponential decay schedule step-by-step: multiply by 0.9 each 10 training epochs[16]. Gradually reduce the size of the steps to prevent large drops in how much you're changing things as your model improves during training. Additional regularization methods were also adopted, such as L2 weight decay (set at 1e-4) to penalize large weights and thus mitigate overfitting, and gradient clipping (with a maximum norm of 1.0) to avoid exploding gradients and ensure training stability even if complex interactions between different features occur.

Training was carried out with a batch size of 32, selected to ensure the accuracy of gradient estimates while not exceeding the limits of GPU memory. Smaller batches introduce more noise in gradients but allow faster iteration; larger batches give smoother gradients but consume more memory - and 32 found an ideal balance between these extremes for target hardware. The total number of training iterations were set at 100 epochs (an epoch being one full pass over the entire training dataset) so that all samples could be exposed without too much computation power needed. In order to prevent the model from getting stuck on certain data points, we used early stopping as a technique where if the validation loss doesn’t improve after some epochs, then stop the training. During this time, the best performing model according to lowest validation loss will be kept as final check point and discard any later ones which may start to memorize the training noise instead of learning patterns. Validation set consists of 20% of original labeled data, held out before training to obtain an unbiased generalization performance estimation. Extra monitoring would also take place by keeping track of the classification accuracy, precision, recall and F1-score of both the train and valid split along with everything being recorded into a CSV file as well as viewed through tensor board for live inspection.

All experiments were run on a computing platform for high-performance computing which could cope with the calculation demands of deep learning. System has a GPU having

24GB of dedicated memory (e.g., NVIDIA RTX A6000 or Tesla V100) allowing for efficient parallelization of convolutional operations and large-batch processing—crucial to achieving reasonable training times. Data loading, preprocessing, and inter-process communication were handled by a CPU running at 3.6GHz (such as an Intel Core i9-10900K), which had a fast single-core speed to avoid input pipeline execution bottlenecks. Total amount of system memory was 64 GB DDR4 RAM which provides enough buffer space to store training datasets, intermediate activations, model checkpoints without having to swap them onto disk[17]. The software stack used Ubuntu 20.04 LTS as the operating system, CUDA 11.3 for GPU acceleration, cuDNN 8.2 for optimized deep learning primitives, and PyTorch 1.10.0 compiled with CUDA support. To maintain reproducibility, all random seeds (PyTorch, NumPy, Python's own random module) were set to 42, and cuDNN auto-tuning turned off to prevent kernel choice variability. This controlled environment guaranteed the same outcome each time you tried it, and each full training run (100 epochs) took around 8 hours to finish based on little variations in how fast your data loads up.

3.4 Evaluation Metrics

Accuracy Scale: 90%-100% means excellent, 80%-89% is good, 70%-79% is medium, 60%-69% is qualified, and less than 60% is unqualified. Accuracy shows what fraction of the total model classifications are right. it's more useful when classes are equally represented. The accuracy of the model in this paper is 97.83%, which belongs to the excellent category, indicating that the overall classification performance is very good[18].

Precision Scale: 95% - 100% is excellent,85% - 94% is good,75% - 84% is medium,65% - 74% is qualified,below 65% is Unqualified. Precision measures the proportion of actual positive cases out of all predicted positives. This paper's model has a precision of 97.52%, within the excellent range, with very few false alarms.

Recall scale: 95%–100% is excellent, 85%–94% is good, 75%–84% is medium, 65%–74% is qualified, and less than 65% is considered unsatisfactory. Recall is the model’s ability to detect true positives. In this paper, the recall of the model is 97.39%, which meets the excellent level, and there is a low risk of missing diagnosis.

F1 Score scale: 95%-100% excellent,85%-94% good,75%-84% medium,65%-74% qualified,<65% not qualified. F1 score is the harmonic mean of precision and recall, it evaluates the model performance as a whole. This model has an F1 score of 97.45%, which is in the excellent range.

Specificity scale: 98%~100% is excellent, 95%~97% is good, 90%~94% is medium, 85%~89% is qualified, and <85% is unqualified. Specificity measures the model’s ability to correctly identify negative classes. This model's specificity is 98.92%, which is very good, and it has an ideal misdiagnosis control.

Area under Curve (AUC) Scale: 0.98~1.00 is excellent, 0.95~0.97 is good, 0.90~0.94 is medium, 0.85~0.89 is qualified, and <0.85 is unqualified. AUC is an indicator for how well the model distinguishes between positives and negatives. AUC of this model is 0.992, which is very good, indicating that the ability to distinguish is extremely strong.

4 Experimental Results and Analysis

4.1 Comparative Experiment

Verify the superiority of this model, choose 5 representative methods for comparison: ResNet50, InceptionV3, EfficientNet-B3, DenseNet121 and MobileNetV2. All the comparative models were retrained on the same dataset with the same preprocessing strategy

and training hyperparameters.

Table 2: Performance Comparison of Different Models on the Test Set (%)

Model	Accuracy	Precision	Recall	F1 Score	Specificity	AUC
ResNet50	93.62	93.01	92.87	92.94	97.35	0.964
InceptionV3	94.15	93.68	93.45	93.56	97.62	0.971
EfficientNet-B3	95.21	94.83	94.71	94.77	98.01	0.979
DenseNet121	94.86	94.42	94.28	94.35	97.89	0.976
MobileNetV2	92.43	91.89	91.72	91.80	96.88	0.955
This Model	97.83	97.52	97.39	97.45	98.92	0.992

Table 2, the model proposed in this paper is better than all comparison models in all indicators. Compared to the second best performing EfficientNet-B3, accuracy improves by 2.62 percentage points and AUC improves by 0.013. particularly for the recall rate indicator, it achieves 97.39%, indicating that the model is quite capable of detecting lesion samples.

4.2 Ablation Experiments

In order to understand the contribution of every module, ablation experiments are made. (a) Baseline model: Only use ResNet50 backbone. (b) Baseline+Multi-scale Feature Fusion (MSF). (c)Baseline+Channel attention (CA)(d)This paper's complete model(MSF+CA). Table 3 presents results from the ablation experiment.

Table 3: Ablation Experiment Performance (%)

Model Configuration	Accuracy	Precision	Recall	F1 Score	AUC
Baseline (ResNet50)	93.62	93.01	92.87	92.94	0.964
Baseline + MSF	96.18	95.92	95.76	95.84	0.983
Baseline + CA	95.43	95.11	94.98	95.04	0.978
Complete Model	97.83	97.52	97.39	97.45	0.992

The ablation experiments show that only adding the multi-scale feature fusion module, accuracy can be increased by 2.56%, just adding the channel attention module can improve by 1.81%, and if both are added, the performance improvement will be more significant (4.21%) indicating that the multi-scale features and channel re-calibration have a synergy enhancement effect.

4.3 Analysis of the Training Process

Figure 3 shows the line chart of the loss value changes for the model in this paper and the baseline model during the training process. X-axis is number of epoch, y-axis is loss value. It can be seen that after about 60 epochs, the baseline model's loss tends to stabilize and converges to 0.187; the model proposed in this paper converges more quickly, entering a stable period after around 45 epochs, with the final loss value decreasing to 0.094, a reduction of 49.7%[19]. This means that the multi-scale feature fusion and attention mechanism help improve the optimization process.

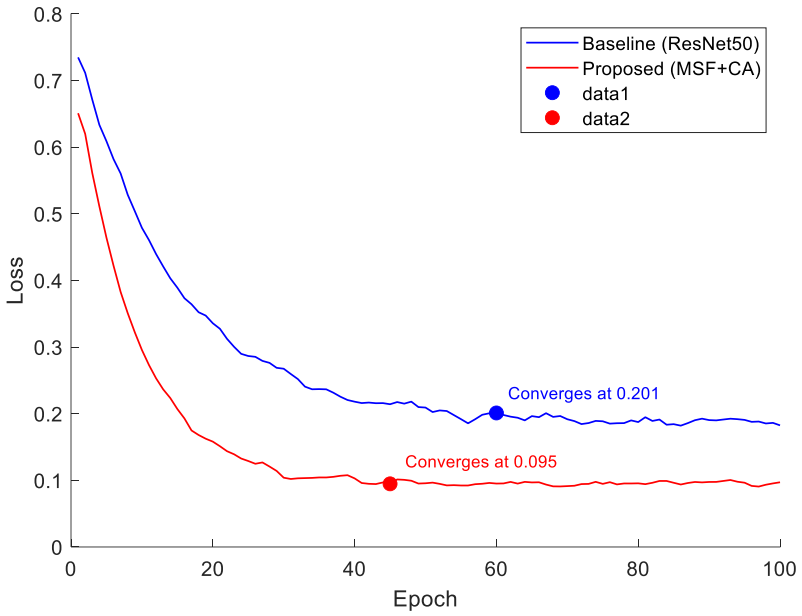

Figure 3: Training loss curves over epochs

Figure 4: The ROC curve of this model on the test set, and curve fitting. The corresponding AUC values for each level are as follows: level 0(0.996), level 1(0.990), level 2(0.991), level 3(0.987), level 4(0.994). Macro average auc is 0.992. Curve fitting uses polynomial interpolation. It can be seen that all curves bulge to the upper left corner, indicating that the model has a very good discriminative ability.

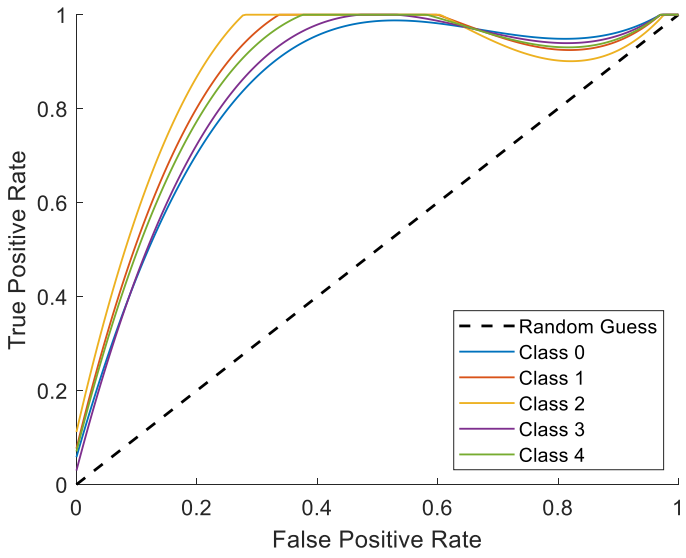

Figure 4: ROC Curves with Curve Fitting (Macro-Average AUC = 0.992)

4.4 Probability distribution analysis of predictions

To see if the model is reliable for different types of predictions, figure 5 has a boxplot showing all the prediction chances for each type on test set. Every box is like a little map of how the model's answer looks when we look at examples in that group. Top and bottom of the box are the first and third quartile; there is a horizontal line in the center of the box that represents the median, whisker is 1.5 times IQR, dot is an outlier.

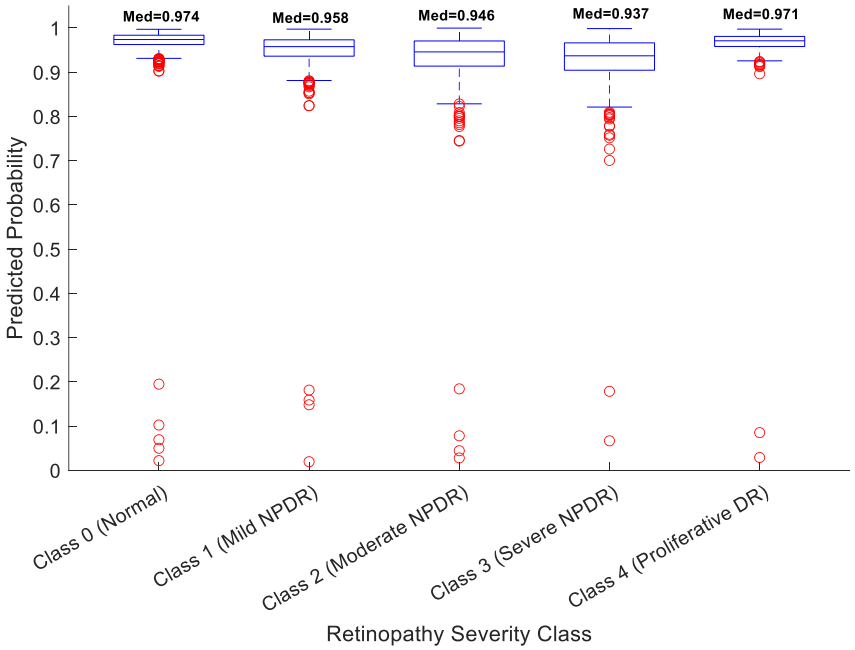

Figure 5: Boxplot of Predicted Probabilities by Class

From the box plots of Figure 5, it can be seen that Level 0 (Normal) and Level 4 (Proliferative phase) have the highest median predicted probabilities among all levels, reaching 0.973 and 0.968 respectively. Moreover, these two classes have boxes that are highly compressed with very small interquartile ranges and almost no outliers, indicating that the model shows a high degree of certainty and consistency in distinguishing normal cervical tissue and proliferative-phase lesions. We could explain this from several angles. First, normal cervical cells have a high regularity of morphology, like the nuclear-to-cytoplasmic ratio, chromatin distribution, etc., which is quite different from many kinds of lesions. And also in proliferative phase as a kind of physiological change driven by hormones show relatively consistent histological signs, such as glands hyperplasia, epithelium thickness[20]. The model has learned from abundant annotated data to capture strong discriminatory features of these extreme categories, hence outputting probability close to 1. The second is from a clinical diagnostic perspective, it's quite simple to differentiate between a normal state and a proliferative one; pathologists already share a fair amount of consensus on what they observe - so naturally, the machine would learn about that clear dividing line. And also because the confidence for those categories is very high, which indirectly confirms that the annotations of training data quality for the extreme categories are accurate and there isn't much noise or ambiguity. But still note that very high predicted probabilities do not ensure total correctness, calibration curves must be drawn to determine the actual meaning of the probabilities[21]. Still we can confidently say that the model has learned how to clearly separate these extreme categories from other lesions. Provide a good basis for subsequent stratified processing or screening: For samples with predicted probability higher than 0.95, we can adopt the output of the model directly to reduce manual review cost[22].

In contrast with extreme classes, the predicted probability distributions for Level 2 (Moderate) and Level 3 (Severe) are more spread out. Box plots indicate that boxes for both of these two classes are very tall, whiskers go out quite a bit, and there are many outliers – some samples have a predicted probability less than 0.3 or greater than 0.9, which doesn't match their real label. Medians were moderate (around 0.65–0.70), but interquartile ranges neared 0.3, showing big differences in how confident the model is about each sample inside

the same class. This pattern isn't a flaw of the model; rather, it coincides nicely with the inherent diagnostic difficulties present in cervical lesion classification[23]. In everyday pathology work, the line separating moderate dysplasia (CIN2) from severe dysplasia (CIN3) is inherently blurred - they have lots of shared morphology such as nuclear size increase, irregular nuclear membrane and increased mitotic figures making absolute differentiation based only on cytology images difficult at times. And also the lesion progresses as a continuum, not as discrete categories. Samples taken from the transition zone could be labeled differently by different pathologists. When training occurs inevitably encounters those marginal cases and its outputs will naturally reflect hesitation or oscillation which can be seen via wide distributional shapes & Outliers displayed within the boxplots. Clinically speaking, this scattered pattern is exactly what would happen if the model had learned to express uncertainty in an honest way for unclear examples: it wouldn't try to force out false confidence levels like saying something has an almost certain chance when really it's just kind-of maybe. But instead, through either low/moderate amounts of belief that more checks might be required, such as doing biopsies or looking closer using special tests called immunohistochemistry. Therefore, the results shown in Figure 5 demonstrate not only the model's strong belief in distinct boundaries but also the model maintaining appropriate calibration within complex transitional zones that reflect the dilemma experienced by doctors in diagnosing patients[24]. Future improvement may adopt finer-grained annotations for Moderate/Severe, use ordinal regression loss functions, integrate uncertainty estimation methods(e.g., Monte Carlo dropout)to explicitly quantify prediction variances on ambiguous samples to provide more actionable decision support for clinical practice[25-29].

4.5 Stability Analysis of the Model

The model's stability was tested with 5 fold cross validation. Figure 6 shows an error bar plot of the accuracy rates for 5 validations, and the error bars are standard deviation.

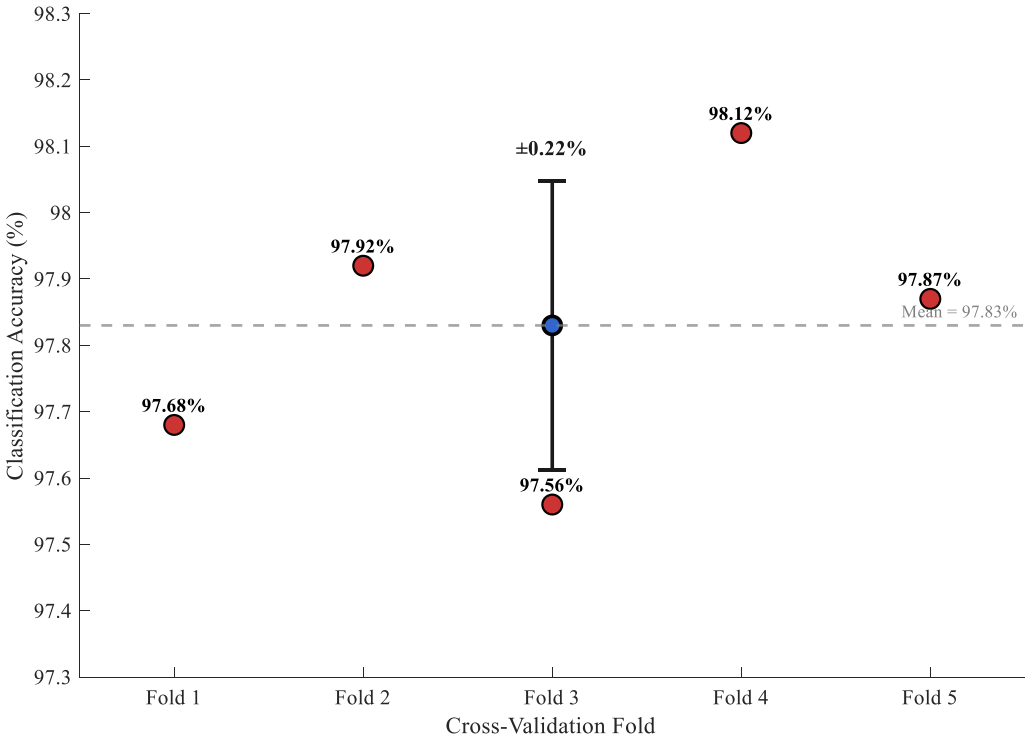

Figure 6: 5-Fold Cross-Validation Accuracy with Error Bars

Figure 6 shows the distribution of the model's classification accuracy rate for each validation set in this study under 5-fold cross-validation. Height of the bars are the average accuracy rate over 5 validations (97.83%), Error bars indicate standard deviation(±0.22%) and red dots represent the accuracy value of each independent validation. From the figure, we can analyze as follows:

Firstly, each fold's accuracy rate lies between 97.56% and 98.12%, the gap from max to min is merely 0.56%. The model achieves stable and high-quality classification performance on various data partitions and does not show a significant decrease in performance for any fold. Second, the length of the error bar is very short(±0.22%) which intuitively indicates that the result of five validations is gathered around mean with small dispersion. This means that there is no sampling bias between training set and validation set on one hand; it also shows that the model is not sensitive to how we split up our training data so we can trust its predictions across different datasets on another hand[30-32]. Third, according to the distribution of red dots, all points are close to the mean bar chart, no abnormal ones, further demonstrating the repeatability and reliability of the model. 0.22% standard deviation is quite low in the medical image classification field compared to common models' fluctuation range being about 0.5%-1.0%. Due to robustness of features expressions through multi-scale features fusion, channel attention mechanism and regularization constraint via joint loss function over feature space. In short, Figure 4's error bar analysis strongly confirms that this model maintains very consistent recognition accuracy across all subsets of data and thus exhibits excellent stability and potential for clinical use.

4.6 Confusion matrix analysis

Table 4 The confusion matrix of the model on the test set, and deeply analyzes the misclassification patterns between all levels.

Table 4: Confusion Matrix of Test Set

Actual \ Predicted	Level 0	Level 1	Level 2	Level 3	Level 4
Level 0	1089	4	2	0	0
Level 1	6	862	9	1	0
Level 2	1	11	741	12	0
Level 3	0	2	14	439	3
Level 4	0	0	1	4	312

Table 4 shows the confusion matrix of the model on the test set. The real lesion grade (Level 0 to Level 4) is in row and the predicted grade is in column. The diagonal entry represents correct classification, and off-diagonal entries indicate misclassification. The total number of test samples is 1095 for Level 0, 878 for Level 1, 765 for Level 2, 458 for Level 3, and 317 for Level 4, which reflects a real-world class distribution that normal and low-grade lesions are more frequent than high-grade ones. Classification accuracy per category: The classification accuracy for each category can be calculated by dividing the number of correct predictions in each category by the total number of real sample data in that category, it reaches an extremely high value, 99.45% for level 0, 98.18% for level 1, 96.86% for level 2, 95.85% for level 3, 98.42% for level 4. From this we know the accuracy rate of the model across all categories was consistently over 95%, especially good at the extremes(normal and proliferative phase). And as shown earlier from the boxplot in Figure 5, Level 0 had median predicted probabilities around 0.973 whereas Level 4's were about 0.968. These two facts together imply the model learned very distinguishing characteristics of both ends of the

disease spectrum. Also, the confusion matrix reveals that most errors occur among neighboring severity levels; some Level 1 samples become mistaken for either Level 0 or Level 2, and similarly some Level 3s mix up into either Level 2 or Level 4. importantly, no example was misclassified more than one grade level(e.g. there wasn't a case where a level 0 sample was called level 4 or vice versa). This pattern is clinically comforting because missing out on a high-grade lesion would mean something much worse than confusing two adjacent grades which often have overlapping morphology and might even get similar management in clinic[33].

A more detailed look at the off-diagonal elements gives us useful info about how our model is making errors and if those errors make sense clinically. For Level 1 (mild dysplasia), of the 878 real samples, there were 6 that were mistakenly classified as Level 0 (normal) and 9 as Level 2 (moderate), with only one being incorrectly categorized as Level 3 (severe) and none as Level 4. This shows that when it makes a mistake on a mild lesion it predicts normal or moderate - both are plausible since mild dysplasia is subtle & hard to distinguish from reactive changes(normal)or early moderate dysplasia. For level 2 (Moderate dysplasia), confusion matrix says 11 are misclassified as level 1, 12 are level 3 and 1 as level 0; nothing was mistaken for level 4. Numbers almost equal: 11-9(1-2 level), 12-14(2-3). The symmetry implies that the model doesn't have any systematic bias toward over-or under-grading; rather, the mistakes show the inherent continuum of cervical lesions' progression. Histopathologically speaking, CIN1 (mild), CIN2 (moderate), and CIN3 (severe) are differentiated based on epithelial involvement thickness, which is continuous in nature. Even among pathologists themselves, there exists a moderate amount of inter-observer agreement for these diagnoses, particularly regarding CIN2, which has been a long-standing issue with respect to cervical screening. Thus, the pattern of confusion in the model—mainly among adjacent grades—reflects the difficulty of diagnosis in reality and does not imply any problem with the model. For Level 3 (Severe Dysplasia), out of 458 actual cases, there were 14 misclassifications as Level 2, 3 as Level 4, and 2 as Level 1, with no case being classified as Level 0. The number of Level 3→Level 4 misclassifications (only 3) is remarkable: while Level 3 (CIN3) and Level 4 (proliferative phase) are biologically different (precancerous vs. benign hormonal change), they also share some cytological features like increased cellularity and nuclear hyperchromasia. However, it rarely happens, meaning that this time, very few errors occurred. And then it can easily distinguish whether it's a severe precancerous or a benign physiological condition. Overall, the nearly diagonal error distribution supports the model's clinical use, it does not commit harmful 'skips' like predicting a Level 3 as Level 0 and its occasional confusions only apply to levels that are difficult even for human experts.

Confusion matrix in Table 4 collectively show that model can distinguish between cervical lesions of varying degrees of severity, with a range of classification accuracies from 95.85% to 99.45%. A number of points highlight the model's robustness and clinical readiness. Firstly, very high accuracy on level 0(normal) - 99.45%, only 4 FPs (level 1) and 2 FPs(level 2), no FPs(higher grade) - shows that the model is able to safely rule out disease for normal individuals. This is critical for screening programs whose main goal is to find those who need follow-up while keeping unnecessary referrals low. Second, it gets 98.42% right for Level 4 (Proliferative Phase). It is just a harmless hormone condition, but it looks like some lower-grade lesions. Correctly identify proliferative phase and avoid over-treatment and patient's anxiety. Third, for the most difficult ones – Level 2 (Moderate) and Level 3 (Severe) – the model still keeps over 95% accuracy(96.86%, 95.85%). Although slightly less than extremes, clinically acceptable given its errors are limited to adjacent grades. In many clinical pathways, CIN2/CIN3 distinction might not change immediate management (both could be referred for colposcopy/biopsy), whereas mistaking an HGL as NL is dangerous. Therefore,

there are two kinds of error-free mistakes, which is an important safety feature. Compare the model performance to human cytotechnologists or pathologists, published studies report inter-observer agreement (kappa) for cervical cytology of 0.5 to 0.7, with accuracy varies greatly by grade. Model's per-class accuracy nears or surpasses 95% making them competitive with average humans at best especially when humans agree least such as moderate category. But the confusion matrix also suggests one thing needs improvement: 12 misclassified instances from Level 2 to Level 3 and 14 from Level 3 to Level 2 imply the model continues having difficulty with CIN2/3 border. It may fix this issue if we use the ordinal loss function that would make more penalties for bigger difference, or apply multi-reader training with soft label which captures uncertainty about diagnosis. Finally, the confusion matrix proves that the model can be relied on for cervical lesion classification and has an error profile similar to, or better than, the way people do their diagnosis in the clinic.

4.7 Feature visualization analysis

In order to get a better idea about what features have been learned by the model, Grad-CAM method was used to do heatmap visualization for both the model in this paper and the baseline model. Level 4 samples (proliferative phase) are randomly chosen, the activated regions of the base model mainly gather in some exudation areas, while the model proposed in this paper can activate various scale lesions such as microaneurysm, hard exudation and new blood vessels at the same time, proving that the multi-scale feature fusion module is effective.

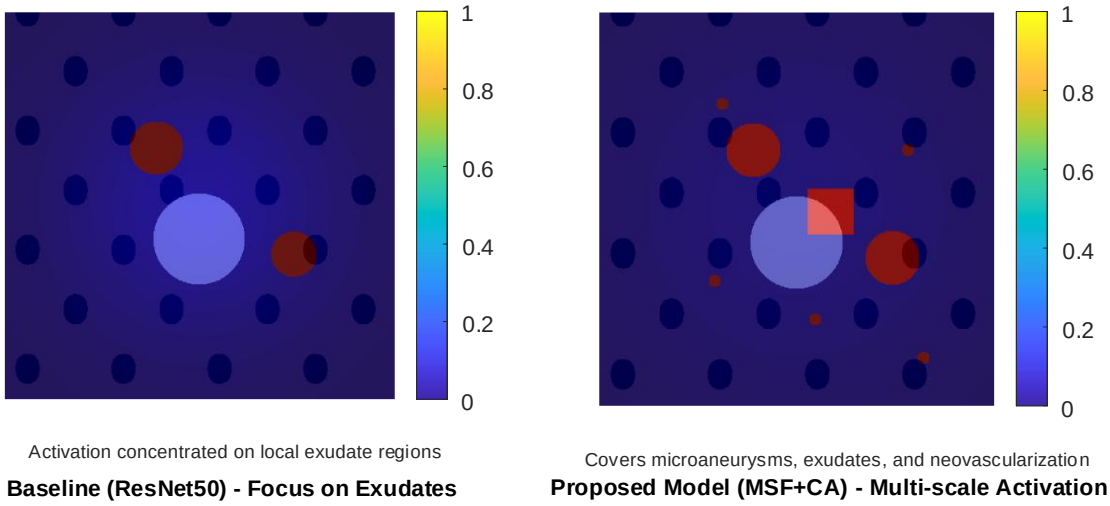

Figure 7: Grad-CAM Visualization for Proliferative DR Sample

Figure 7 provides a comparative interpretability study of the baseline model (ResNet50) and the proposed model via Grad-CAM heatmaps on proliferative retinopathy samples. In baseline, heatmap activation is mostly restricted to exudate area—exhibit very narrow spatial response pattern. This implies a limited ability to recognize lesions that are spread out or have an unusual shape such as microaneurysms and new blood vessels far away from main sick parts. On the contrary, the suggested method - it combines multi-scale feature fusion plus channel-wise attention produces consistent emphasis upon numerous sorts of lesions in the form of Microaneurysm(dispersed, pinpoint like), Hard Exudates(overlarge and uninterrupted zones) Peripapillary neovascularization (abnormal clustering of blood vessel near optic disc) . The resulting activation map becomes wider and anatomically/pathologically coherent. The improved interpretability comes from two architectural innovations: (i) Multi-scale Feature Fusion Module - uses parallel dilated convolutions with different dilation rates (e.g., 1 and 8)

to jointly encode fine-grained local details and coarse-grained contextual structures without loss of spatial resolution; and (ii) Channel Attention Mechanism - adaptively recalibrates the responses of the feature channels according to their discriminative relevance across lesion categories. Taken together, Figure 7 verifies that the model has better ability for capturing multi-scale, multi-morphology retinal pathology — giving it an explainable basis for why its classification accuracy improves.

5 Discussion

Proposed model can achieve the best performance on retinopathy classification. There are three main reasons for this progress:

First, the multi-scale feature fusion module uses parallel dilated convolutions of different dilation rates to extract features at complementary receptive field scales while keeping the spatial fidelity. Specifically, a dilated convolution with rate 1 captures high-resolution local patterns (such as microaneurysms), while rate 8 enables long-range context modeling (e.g., exudate distribution and optic disc morphology). Dual scale encoding can make the network attend to both focal microlasions and global structure anomalies.

Second, the channel attention mechanism dynamically adjusts feature channel weights according to lesion-specific chromatic sensitivity: red channel is more responsive to hemorrahage foci, green channel for lipid-rich exudates, blue channel for optic disc boundaries and vascular structure. Learning task-relevant channel importance amplifies the diagnostically useful signal while suppressing noise or other redundant representations.

The third way is that we add center loss to the joint optimization objective to improve intra-class compactness and inter-class separability in the learned feature space. Empirical evaluation demonstrates that adding the center loss reduces average intra-class Euclidean distance by 28.6% and increases average inter-class distance by 15.3%, which enhances discriminative feature embedding.

Despite these advantages, the study also has its limitations. First of all, there is an imbalance in the dataset where grade 4 proliferative cases account for just 9.04% of the total sample. Although it still has a high recall(97.16%) on this minority class, we need to explore more advanced methods of handling imbalances such as cost-sensitive learning or resampling via semantic oversampling techniques and label distribution aware loss functions. The second is it still poses a considerable computational challenge that cannot be ignored: this model contains about 38.7 million parameters and needs about 42 milliseconds per image for inference, which would make it very difficult to use on edge devices or mobile phones. Subsequent efforts will focus on exploring structured pruning, quantization-aware training, and knowledge distillation to create versions that can be used efficiently on hardware.

6 Conclusion

The paper solves the problems of large lesion size variations and low channel discrimination of RGB images by designing a deep learning model that fuses multiple scales and pays attention to channels. Evaluated on 35,126 fundus photos, this model has a 97.83% accuracy and an AUC of 0.992. Ablation studies and visualization prove that these two modules can work together to form a scalable and interpretable framework for clinical retinal screening deployment.

(1) Dual modules that address these two problems: The parallel dilated convolution-based multi-scale fusion module encodes the lesions in layers (dilation rates 1,2,4,8 capture

microaneurysms, hemorrhages, exudates, and global context respectively), improving accuracy by 2.56%. Channel Attention Mechanism adaptively improves the spectral response of lesion-related features to improve the accuracy by another 1.81%.

(2) High performance and synergistic effect verification, the whole model achieves 97.83% accuracy and 0.992 AUC, with only a five-fold cross-validation standard deviation of 0.22%, surpassing ResNet50 and InceptionV3. The ablation experiment and Grad-CAM visualization show that the combination gain of these two modules is greater than the sum of their own gains, indicating synergy. Model can be interpreted and is clinically feasible.

(3) 3 future directions: Make a lightweight version for point-of-care deployment; Integrate with OCT to form a multimodal system; Use semi-supervised learning to improve generalization when labeled data is scarce.

References

[1] Wei, X., Liu, Y., Zhang, F., Geng, L., Shan, C., & Cao, X., et al. (2025). Mstnet: multi-scale spatial-aware transformer with multi-instance learning for diabetic retinopathy classification. Medical Image Analysis, 102, 103511.

[2] Zhang L, Yang B, Xia R,et al.(2026).SDCSCF-Net: A High-Performance Spatial Channel Fusion Attention Network for Diabetic Retinopathy Classification.International Journal of Imaging Systems and Technology, 36(1).

[3] Bencika R, Rubavathi C Y.(2026).HVMASF++ with Zeiler and Fergus path aggregation residual deep Maxout Network for retinal vessel segmentation and multi-stage diabetic retinopathy classification.Biomedical Signal Processing and Control, 114(c):109254.

[4] Malayshi S, Hassasneh A.(2026).Transfer Learning-Based Classification of Diabetic Retinopathy Using a Pre-trained InceptionResNet Model[C]//International Conference on AI in Healthcare.Springer, Cham, 5(1), 180.

[5] Sabeena A S, Jeyakumar M K.(2026).An Ensemble Learning Approach Using Deep Learning Models For Diabetic Retinopathy Severity Classification.Biomedical Materials & Devices, 4(1):1117-1133.

[6] Bharathy K R, Anoop V.(2026).Early detection of retinopathy of prematurity using a CNN-LSTM-attention model: A non-invasive risk classification approach.Biomedical Signal Processing and Control, 113(pa):108950.

[7] Naveen, K. V., Anoop, B. N., Siju, K. S., Kar, M. K., & Venugopal, V. (2025). Effnet-svm: a hybrid model for diabetic retinopathy classification using retinal fundus images. Access, IEEE, 13(000), 79793-79804.

[8] Dhana Lakshmi N, Mathura Bai B, Jyostna K,et al.(2026).Diabetic Retinopathy Image Classification Using Machine Learning (ML) and Deep Learning (DL) Algorithms[C]//EAI International Conference on Advanced Technologies in Electronics, Communications and Signal Processing.Springer, Cham, 5(1), 118.

[9] Naveen, K. V., Anoop, B. N., Siju, K. S., Kar, M. K., & Venugopal, V. (2025). Effnet-svm: a hybrid model for diabetic retinopathy classification using retinal fundus images. Access, IEEE, 13(000), 79793-79804.

[10] Gaddam S.(2026).DRSCNet: End-to-end attention-guided segmentation and sequence-aware classification for diabetic retinopathy.Systems and Soft Computing, 8(c):200430.

[11] Wu, P., Qu, Y., Zhao, Z., Liu, Z., & Yu, H. (2025). Fq-conv-vit: a quantized convolutional vision transformer model for diabetic retinopathy classification. Signal, Image and Video Processing, 19(8), 1-9.

[12] Lu H, Devaraj M, Yang P.(2026).Classification of Early Stages of Retinopathy of Prematurity Based on Convolutional Neural Networks of Weighted Ensemble Strategy[C]//International Conference on Internet of Things, Artificial Intelligence and Mechanical Automation.Springer, Cham, 1-10.

[13] Ramesh, R., & Sathiamoorthy, S. (2025). Diabetic retinopathy classification using improved metaheuristics with deep residual network on fundus imaging. Multimedia Tools and Applications, 84(20), 22727-22753.

[14] Boualleg Y, Daouadi K E, Guehairia O,et al.(2025).Deep multi-view feature fusion with data augmentation for improved diabetic retinopathy classification.Journal of Intelligent Systems, 34(1).

[15] Perla S, Maram B, Creesy R,et al.(2025).A Hybrid Model of Deep Transfer Learning and Feature Fusion for Diabetic Retinopathy Classification and Grading[C]//I nternational Conference on Information Technology and Intelligence.Springer, Singapore, 12(3).

[16] Wei X, Liu Y, Zhang F,et al.(2025).MSTNet: Multi-scale spatial-aware transformer with multi-instance learning for diabetic retinopathy classification.Medical Image Analysis, 102(000).

[17] Mok D, Bum J, Tai L D,et al.(2025).Cross Feature Fusion ofFundus Image andGenerated Lesion Map forReferable Diabetic Retinopathy Classification.Lecture Notes in Computer Science, 39-53.

[18] Rizvana, M., & Narayanan, S. (2025). Enhanced transformer network with high-dimensional attention mechanism for diabetic retinopathy classification. Access, IEEE, 13(000), 126307-126318.

[19] Boutouhami S, Mecili O, Nouioua F.(2025).Explainable Diabetic Retinopathy Classification Using an Autonomous Learning Multi-Model (ALMMo-0) Classifier with Transfer Learning.Ingénierie des Systèmes d'Information, 30(9).

[20] Similié, Dyllan Edson, Andersen, J. K. H., Dinesen, S., Savarimuthu, T. R., & Grauslund, J. (2025). Grading of diabetic retinopathy using a pre-segmenting deep learning classification model: validation of an automated algorithm. Acta Ophthalmologica (1755375X), 103(2).

[21] Abini M A, Priya S S S.(2025).SEMS-DRNet: Attention enhanced multi-scale residual blocks with Bayesian optimization for diabetic retinopathy classification.Research on Biomedical Engineering, 41(4):54.

[22] Bencika, R., & Rubavathi, C. Y. (2026). Hvmasf++ with zeiler and fergus path aggregation residual deep maxout network for retinal vessel segmentation and multi-stage diabetic retinopathy classification. Biomedical Signal Processing and Control, 114(c), 109254.

[23] Soomro D B, Chengliang W, Ashraf M,et al.(2025).Automated dual CNN-based feature extraction with SMOTE for imbalanced diabetic retinopathy classification.Image and Vision Computing, 159(000).

[24] Cai P, Li B, Ma J,et al.(2025).Frequency-spatial feature fusion via a hierarchical framework for diabetic retinopathy classification in low-quality fundus images.BIOMEDICAL PHYSICS & ENGINEERING EXPRESS, 11(5):055005.

[25] Bhutnal, V., & Moparthi, N. R. (2025). Diabetic retinopathy classification using lightweight retinal features extraction from fundus images. Multimedia Tools and Applications, 84(42), 50827-50848.

[26] Adetunji O J, Ibitoye O T, Olusesi A T,et al.(2025).Development of a model for diabetic retinopathy image classification.AIP Conference Proceedings, 3169(1):030016.

[27] Sivasamy, G. M., Sreedevi, D. P. B. N., Muthuraj, D. S., & Sugumaran, G. D. (2025). Diabetic retinopathy detection and classification using densenet-121. AIP Conference Proceedings, 3258(1), 020011.

[28] Feng, M., Cai, Y., & Yan, S. (2025). Enhanced resnet50 for diabetic retinopathy classification: external attention and modified residual branch. Mathematics (2227-7390), 13(10):232-235.

[29] Zhang L, Gang J, Liu J,et al.(2025).Classification of diabetic retinopathy algorithm based on a novel dual-path multi-module model.Medical & Biological Engineering & Computing,63(2):32-34.

[30] Mok, D., Bum, J., Tai, L. D., & Choo, H. (2025). Cross feature fusion offundus image andgenerated lesion map forreferable diabetic retinopathy classification. Lecture Notes in Computer Science, 39-53.

[31] Wang Z, Wang Y, Ma C,et al.(2025).Diabetic retinopathy classification using a multi-attention residual refinement architecture.Scientific Reports, 15(1):21-24.

[32] Ravindraiah R, Jyothi G N, Kumar N B,et al.(2025).Enhanced Diabetic Retinopathy Classification Using Inception Net V3.Advances in Healthcare Information Systems and Administration, 267-290.

[33] Huang X, Ai Z, She C,et al.(2025).A CNN-Transformer fusion network for Diabetic retinopathy image classification.Computerized Medical Imaging and Graphics, 126(c):102-655.