scientific reports

OPEN

Transformer attention fusion for fine grained medical image classification

Danyal Badar[1], Junaid Abbas[2], Raed Alsini[3], Tahir Abbas[4], Wang ChengLiang[1]✉ & Ali Daud[5]✉

Fine-grained visual classification is fundamental for medical image applications because it detects minor lesions. Diabetic retinopathy (DR) is a preventable cause of blindness, which requires exact and timely diagnosis to prevent vision damage. The challenges automated DR classification systems face include irregular lesions, uneven distributions between image classes, and inconsistent image quality that reduces diagnostic accuracy during early detection stages. Our solution to these problems includes MSCAS-Net (Multi-Scale Cross and Self-Attention Network), which uses the Swin Transformer as the backbone. It extracts features at three different resolutions (12 × 12, 24 × 24, 48 × 48), allowing it to detect subtle local features and global elements. This model uses self-attention mechanics to improve spatial connections between single scales and cross-attention to automatically match feature patterns across multiple scales, thereby developing a comprehensive information structure. The model becomes better at detecting significant lesions because of its dual mechanism, which focuses on both attention points. MSCAS-Net displays the best performance on APTOS and DDR and IDRID benchmarks by reaching accuracy levels of 93.8%, 89.80% and 86.70%, respectively. Through its algorithm, the model solves problems with imbalanced datasets and inconsistent image quality without needing data augmentation because it learns stable features. MSCAS-Net demonstrates a breakthrough in automated DR diagnostics since it combines high diagnostic precision with interpretable abilities to become an efficient AI-powered clinical decision support system. The presented research demonstrates how fine-grained visual classification methods benefit detecting and treating DR during its early stages.

Keywords Fine-grained visual classification, Multi-scale feature extraction, Attention mechanism, Deep learning, Medical images, Diabetic retinopathy classification

DR is a progressive eye disease that develops due to long-standing diabetes, affecting the blood vessels of the retina. Thus, it is a leading cause of preventable blindness worldwide, including non-proliferative to proliferative retinopathy stages of different severity[1,2]. Since vision loss is prevented by early diagnosis and treatment, automated classification systems are critical to help healthcare professionals diagnose. The classification of DR has several difficulties. The variability in lesion appearances, like microaneurysms, hemorrhages, and exudates, can vary quite a bit between patients, which is one of the significant problems faced. Additionally, comorbidities, including cataracts, can hinder the evaluation of retinal images due to obscuration[3]. There is, in fact, class imbalance in datasets, and DR is in the advanced stages of that condition, which means we have too few of them in the dataset and thus also in the machine learning model training[4].

Moreover, for accurate diagnosis, pictures of high-resolution fundus images are necessary, which require large amounts of computational resources[5]. Since it is necessary to make clinical decisions, the biggest challenge related to the use of automated systems in healthcare is to provide explainability and trustworthiness of the model predictions to healthcare professionals to validate clinical decisions that are based on the model's predictions[6]. Nonetheless, it is crucial to acknowledge the complexities faced when working on DR classification systems, including the competition's large-scale dataset and the need for better data augmentation, deep learning methods, and synthetic oversampling methods to enhance the DR classification models' performance[7].

[1]College of Computer Science, Chongqing University, Chongqing, China. [2]School of Big Data and Software Engineering, Chongqing University, Chongqing, China. [3]Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. [4]Department of Computer Science, TIMES Institute, Multan 60000, Pakistan. [5]Faculty of Resilience, Rabdan Academy, Abu Dhabi, United Arab Emirates. ✉email: wangcl@cqu.edu.cn; alimsdb@gmail.com

However, the general methods for detecting subtle features in medical images suffer from losses in their diagnosis due to anatomical variations, obscured abnormalities (e.g., microaneurysm or faint lesions), or noise, making it not easy to detect early-stage diagnosis. In most deep learning approaches, details are ignored, and only essential patterns are captured. Simultaneous model training is tricky, as different equipment and conditions will inevitably bring about different image qualities and imbalances in the dataset. In addition, clinicians' trust in the predictions is impaired when dealing with minute abnormalities because the predictions lack interpretability.

To deal with these issues, we use attention mechanisms that increase sensitivity by focusing models on important image regions and data generation (e.g., GANs) as well as those that introduce data heterogeneity (i.e., diverse datasets) that can mitigate class imbalance[8]. CNN, together with the transformer, can work robustly and embed global context and local details[9]. In contrast, fine-grained classification techniques based on self-attention mechanisms aim to classify within subclasses by fine features without being limited by the use of single-scale outputs, background interference, and texture similarities[10]. Since these limitations have been identified, advancements in multi-scale feature extraction and improved attention mechanisms are needed to overcome them[11]. These strategies improve the accuracy and confidence of both medical image processing systems and the valuable detection of subtle but clinically essential features in, for example, diabetic retinopathy and oncologic imaging.

To solve these problems, the MSCAS-Net model introduces a multistage feature extraction model that builds the patches of different scales (12 × 12, 24 × 24, 48 × 48) to have both fine and coarse-grained information of DR images. This method solves the feature extraction problem in complex datasets such as APTOS 2019 and DDR using a self-attention mechanism to capture local dependencies and a cross-attention mechanism to integrate global context for effective feature combination at various scales. The model tackles the dataset imbalance for APTOS 2019 and IDRID by focusing on subtle lesions that are often missed in such cases. Cross-attention alleviates the overfitting of the dominant classes and makes a balanced representation across all the DR severity levels. Furthermore, the model naturally has inherent robustness to dataset variations due to the multi-scale architecture of the model and thus requires minimal augmentation. MSCAS-Net learns more invariant features and is less sensitive to variations in orientation, size, and quality by processing images at multiple scales. Self-attention takes care of allowing relevant features that exist in the limited dataset. The model also deals with the variability in image quality by the convolutional layer adjustments, i.e., 1 × 1, 2 × 2, 4 × 4, and the multi-scale processing, which allows the model to cope with the different resolution and quality levels. The proposed two models demonstrate robust performance in low-quality cases by employing cross-attention for maintaining global context and self-attention for attention focused on relevant features for low-quality cases. With such a comprehensive approach, MSCAS-Net is highly effective for DR classification and outperforms state-of-the-art on various datasets.

Therefore, we have performed a detailed study on three data sets, APTOS, DDR, & IDRID, and demonstrated high performance. In conclusion, the paper has the following key contributions:

- Novel MSCAS-Net Architecture: Introduces a Multi-Scale Cross and Self-Attention Network tailored for DR classification, integrating multi-scale feature extraction and advanced attention mechanisms to handle complex medical image analysis challenges like irregular lesions and variable image quality.
- Hierarchical Feature Extraction: Utilizes a Swin Transformer backbone to extract features at three resolutions (12 × 12, 24 × 24, 48 × 48), capturing both fine local details (e.g., microaneurysms) and global retinal structures for effective early-stage DR detection.
- Dual Attention Mechanism: Combines self-attention to enhance local dependencies within each scale and cross-attention to align and fuse features across scales, improving diagnostic precision by capturing hierarchical relationships.
- Robustness to Dataset Issues: Addresses DR dataset challenges like class imbalance and inconsistent image quality without heavy reliance on data augmentation, leveraging its multi-scale and attention mechanisms for stable feature learning.
- High Performance and Interpretability: Achieves top accuracy on benchmark datasets (APTOS: 93.8%, DDR: 89.80%, IDRID: 86.70%) and provides interpretable heatmaps to highlight lesion regions, aiding clinical decision-making and trust in AI diagnostics.
- Efficiency and Generalization: Employs efficient shifted window-based self-attention for reduced computational overhead and demonstrates strong cross-dataset generalization across diverse datasets, making it adaptable for resource-constrained clinical settings.

Related work

Traditional image classification approaches

Convolutional Neural Networks (CNNs) are a type of neural network designed to work on spatial or temporal hierarchical data. They can perform tasks such as image classification, object detection, medical image analysis, etc. Convolutional layers extract spatial features, the dimensions of the space are reduced by pooling layers, and activation functions are used, e.g., ReLU. Solutions to vanishing gradients and scaling have transformed computer vision with the introduction of transformations to scale models and further architectures such as AlexNet[12], VGGNet[13], ResNet[14] and EfficientNet[5]. Both RNNs and transformers can handle sequential data, and transformers excel in language and vision tasks with computational cost. However, beyond CNNs, both RNNs and transformers have limitations in long dependency[3].

Interestingly, however, GANs can produce truly realistic samples[8]Autoencoders can be helpful for dimensionality reduction in anomaly detection, and GNNs can be employed to handle graph data in molecular modeling. Each architecture has strengths and weaknesses, so the architecture selection is problem-dependent on the domain, data type, and computational constraints[5]. The biomedical field is utilizing text mining to extract insights from clinic reports, research papers, and lab tests, focusing on protein-protein interactions and entity-relationship detection, enabling drug discovery and personalized treatment[15]. Integration of IoMT and IoT in healthcare, focusing on security threats and performance enhancement, highlighting limitations in Industry 5.0 and healthcare institutions[16]. Wireless body area networks integrate cloud computing for real-time patient monitoring, but new data privacy and security threats arise. A six-step framework for PPPs privacy and security is provided[17].

Mondal et al. introduced EDLDR, combining DenseNet101 and ResNeXt architectures in an ensemble with GAN-based data augmentation. The method achieved 86.08% accuracy for severity diagnosis and 96.98% for DR identification but faced challenges with data imbalance across categories[7]. These works demonstrate advances in DR classification using attention mechanisms, transfer learning, and ensemble techniques. The cross-disease attention network (CANet) is used to diagnose diabetic retinopathy and macular edema using individual attention mechanisms for feature selection. The model achieved 92.6% accuracy on the Messidor dataset and 65.1% on the IDRiD dataset[6]. Blockchain technology offers a Blockchain-Based Access Control Model (BBACM) for managing authorization rights for accessing patient physiological parameters and PHI, improving access control, security, privacy, scalability, and PHI availability in healthcare data management[18].

Capsule network-based approach, employing CLAHE preprocessing and a sigmoid classifier, resulting in 99.1% accuracy on the Messidor dataset[13]. EfficientNet-B0 with preprocessing to normalize illumination variance, achieving 86.2% accuracy on the APTOS dataset and 84.8% on DDR, though the results were limited by dataset dependency[19]. Similarly, Islam et al. applied CLAHE preprocessing with a transfer learning-based Xception model, attaining 98.36% and 84.36% accuracy for DR identification and severity detection, respectively[20]. Deep learning architectures for analyzing digital social media data, addressing challenges like scalability, heterogeneity, and multimodality, and predicting future trends in social media analytics[21–23].

Recent transformer-based models

Initially developed for natural language processing, transformer-based models are now widely used in medical image processing due to their ability to model long-range dependencies and capture global context via self-attention mechanisms. Vision Transformers (ViT) adapt transformers for image tasks by segmenting images into patches, enabling efficient feature extraction without convolutions. Hybrid models integrating CNNs and transformers combine local and global features, excelling in classification and segmentation tasks despite high computational demands and extensive dataset requirements[24,25]. Zhao et al.[9] introduced CoT-XNet, integrating a contextual transformer with Xception, achieving promising results on DR datasets like DDR, APTOS, and EyePACS. Similarly, Ali Dihin et al.[10] and Yang et al.[26] demonstrated the Swin Transformer's effectiveness in DR grading, emphasizing its efficient attention mechanisms. Yan Y[27]. proposed AlexViT, blending AlexNet with ViT for DR classification, achieving 88.23% accuracy and notable efficiency. Attention mechanisms enhance feature extraction by focusing on critical regions in images, improving interpretability, and highlighting pathological areas. Madarapu et al.[11] developed a deep integrative model for DR classification combining residual blocks, channel-spatial attention mechanisms, and non-local blocks for robust feature representation. Furthermore, Li et al.[28] utilized ConvNeXt-base with attention mechanisms for DR detection, achieving high diagnostic accuracy and interpretability and demonstrating significant advancements in attention-based medical imaging approaches.

Fine-grained classification

Fine-graining in image processing[29] involves the detailed categorization of objects within a broader category, often requiring advanced methodologies due to the subtle differences between classes. Techniques such as attention mechanisms and deep learning architectures have enhanced fine-grained recognition. The use of region grouping for interpretable and accurate recognition[30] focuses on identifying the most discriminative regions within images. Self-supervised structure modeling, as seen in Look-into-Object[31] enables object recognition by focusing on internal structures. Deep Convolutional Neural Networks (DCNN) in diagnosing and classifying lung cancer using medical imaging, highlighting their significant contributions to early detection and treatment, focusing on 2015–2024[32].

Dual cross-attention mechanisms[33] and recurrent attention with multi-scale transformers[34] offer robust solutions for refining classification through interrelated feature extraction. Salient mask-guided vision transformers[35] and channel interaction networks[36] emphasize interaction and region-specific attention for improved accuracy. Moreover, counterfactual attention learning[37] and Gaussian mixture models[38] address weak supervision, enabling enhanced discrimination without exhaustive labeling. These approaches collectively advance fine-grained image recognition, ensuring precise categorization.

Table 1 provides a comprehensive comparison of state-of-the-art CNN-based methods and techniques, and Table 2 highlights advancements in Transformer-based approaches, showcasing their effectiveness in medical image analysis, specifically for DR detection and classification.

Methodology

The diagram shows that MSCAS-Net (Multi-Scale Cross and Self-Attention network) is made to extract more features from an input image for classification tasks ranging from convolution processing to advanced attention mechanisms. So, we split up the input image into specific sizes of patches using a Swin Transformer. However, this transformer splits the image at 3 scales (12 × 12 × 1024, 24 × 24 × 512, 48 × 48 × 256) to capture features of different granularities. This multi-scale feature extraction provides fine and coarse details to be used further. Complementary information is given in each scale to improve feature richness. The features are then processed with three convolutions: 1 × 1, 3 × 3, and 4 × 3, and each is a fixed size of 1 × 1, 3 × 3, and 4 × 3, respectively — as a way of adjusting the spatial dimension of the features whilst improving their representation and maintaining

Approaches	Dataset	DR grading (classes)	Accuracy (%)
Mondal et al.[7]	APTOS	5	86.08
Zhao et al.[9]	DDR, APTOS, EyePACS	5	83.10, 84.18, 84.10
Vijayan et al.[19]	APTOS, DDR	5	86.20, 84.80
Islam et al.[20]	APTOS	2, 5	98.36, 84.36
Oulhadj et al.[39]	APTOS	5	85.28
Oulhadj et al.[40]	APTOS	5	86.54
Bodapati et al.[41]	APTOS, IDRiD	5	84.17, 63.24
Fan et al.[42]	APTOS	5	85.32
Sugeno et al.[43]	APTOS	5	84.20
Shaik and Cherukuri[44]	APTOS, IDRiD	5	85.54, 66.41
Al-Antary and Arafa[45]	APTOS, EyePACS	2, 5	98.10, M:84.60 87.50, M:79.90
Abbasi et al.[46]	Messidor, EyePACS	5	82.32, 76.84
Badar[22]	APTOS	3	92.73

Table 1. CNN-based state-of-the-art DR detection and Classification.

Study	Model	Dataset	Classes	Accuracy
Ali Dihin[10]	Dhin Window Transformer	APTOS	5	85.3%
Yaoming Yang[26]	Swin Transformer	APTOS	5	85.3%,
Yan[27]	AlexVit	APTOS	5	88.23%
Mohammed Oulhadj[40]	Vision Transformer	APTOS	5	88.18%
Dihin[10]	Wavelet-Attention Swin	APTOS	5	86%

Table 2. Transformer-based state-of-the-art DR detection and Classification.

scale consistency. With this step, the model is able to process and refine the extracted features better so that the most critical spatial relations within the image are preserved.

The features are then processed through self-attention blocks, which will help the model focus on key areas of the image by learning internal dependency within the features. The self-attention mechanism guarantees that the model does not have to rely on the global pattern of the image but can capture locally relevant patterns in the feature maps. Via a cross-attention block, the output from the self-attention layers is merged. By utilizing the integrated local and global contexts through the cross-attention block, the model will enhance its ability to understand the overall structure and the crucial parts in the image. Instead of cross-attention, it takes the queries, keys, and values from the outputs of self-attention and then focuses on the most important and relevant parts between the different scales. Finally, the result of the cross-attention block is combined into a single feature vector that contains the local and global feature information in one comprehensive representation. The image is fed to the classification layer with the extracted and processed feature and is then classified according to this vector. It is a good model to improve classification performance while exploiting a comprehensive combination of multi-scale feature extraction, convolutional processing, self-attention, and cross-attention mechanisms that are also highly efficient for image classification tasks.

Multi-scale feature extraction

The input image $I \in R^{H \times W \times C}$ be passed through the Swin Transformer backbone f_{Swin} for multi-scale feature extraction. The image is split into patches at three distinct scales, resulting in feature maps at three different levels:

$$F_{\text{level1}} = f_{\text{Swin}}\,(I, \text{scale} = 12 \times 12) \in R^{H_1 \times W_1 \times C_1} \tag{1}$$

$$F_{\text{level2}} = f_{\text{Swin}}\,(I, \text{scale} = 24 \times 24) \in R^{H_2 \times W_2 \times C_2} \tag{2}$$

$$F_{\text{level3}} = f_{\text{Swin}}\,(I, \text{scale} = 48 \times 48) \in R^{H_3 \times W_3 \times C_3} \tag{3}$$

.

Where F_{level1} represents global features extracted using large patches of $12 \times 12 \times 1024$, F_{level2} captures mid-level features using $24 \times 24 \times 512$ patches and F_{level3} extracts fine-grained local details using smaller $48 \times 48 \times 256$ patches.

After that, each feature map is convolved with differing kernel sizes via convolutional layers that would preserve the basic characteristics while maintaining the same spatial size for feature levels.:

$$F_{\text{level1, conv}} = \text{Conv}_{1 \times 1}\,(F_{\text{level1}}) \in R^{H_1 \times W_1 \times C_1} \tag{4}$$

$$F_{\text{level2, conv}} = \text{Conv}_{2\times\, 2}\left(F_{\text{level2}}\right) \in\ R^{H_2\times\, W_2\times\, C_2} \tag{5}$$

$$F_{\text{level3, conv}} = \text{Conv}_{4\times\, 4}\left(F_{\text{level3}}\right) \in\ R^{H_3\times\, W_3\times\, C_3} \tag{6}$$

.

Simplifying the integration and comparison of features is a matter of spatial size uniformity, and thus reduces computational time. Once the spatial dimensions are aligned, appropriate fusion strategies can be employed, e.g., concatenation or attention-based, to simultaneously obtain global context and local context for the model. After applying the convolution operations with different kernels, all feature maps (global, mid-level, and local) are transformed to a consistent size of $12\times\ 12$ but still retain their unique characteristics.

$$F_{\text{level1, conv}} \in\ R^{12\times\ 12\times\ 1024} \tag{7}$$

$$F_{\text{level2, conv}} \in\ R^{12\times\ 12\times\ 1024} \tag{8}$$

$$F_{\text{level3, conv}} \in\ R^{12\times\ 12\times\ 1024} \tag{9}$$

.

Once the features are transformed to a uniform spatial size of $12\times\ 12$ the feature maps are passed to subsequent processing layers for further feature refinement.

Figure 1 illustrates the Architecture of MSCAS-Net (Multi-Scale Cross and Self-Attention Network) for diabetic retinopathy classification. The model divides input images into multi-scale patches for hierarchical feature extraction. Self-attention refines local dependencies within each scale, while cross-attention integrates global context by aligning features across scales. Convolutional layers adjust spatial dimensions, and the unified feature vector combines local and global information for final classification.

Hierarchical feature refinement via self-attention

After transforming the multi-scale feature maps to a consistent spatial size $12\times\ 12$ while retaining their unique characteristics, the feature maps $\boldsymbol{F}_{\text{level1, conv}}$, $\boldsymbol{F}_{\text{level2, conv}}$, and $\boldsymbol{F}_{\text{level3, conv}}$ are processed independently through self-attention layers to refine their discriminative capabilities. Each feature map is projected into query $\boldsymbol{Q_i}$, key $\boldsymbol{K_i}$, and value $\boldsymbol{V_i}$ embeddings using learnable weight matrices:

$$Q_i = W_{q,i}\cdot\ F_{leveli,conv} \tag{10}$$

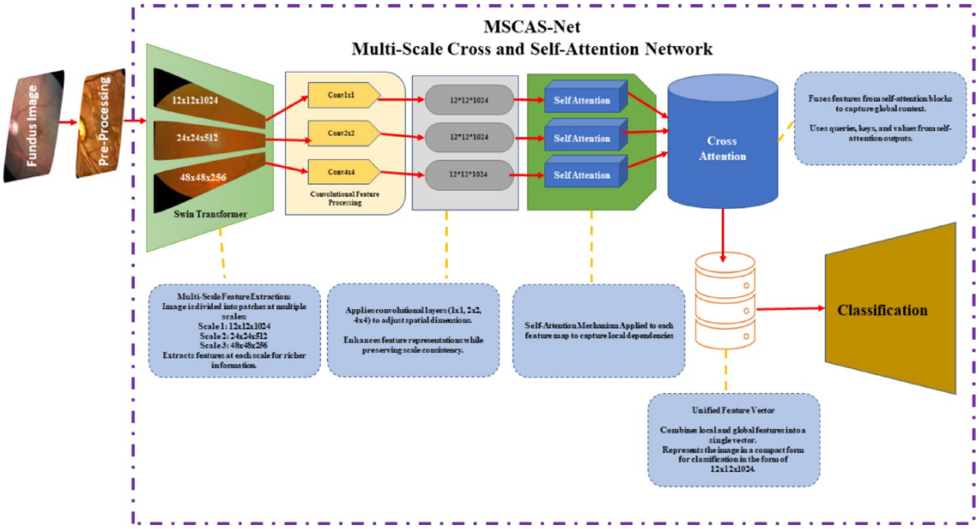
$$K_i = W_{k,i}\cdot\ F_{leveli,conv} \tag{11}$$

$$V_i = W_{v,i}\cdot\ F_{leveli,conv} \tag{12}$$

.

Where $i\ \in\ \{1,\ 2,\ 3\}$ denotes the level index, and $W_{q,i}$, $W_{k,i}$, and $W_{v,i}$ are learnable matrices. These embeddings enable the model to compute level-specific attention using the scaled dot-product mechanism:

$$A_i = \text{Softmax}\left(\frac{Q_i\cdot\ K_i^{\top}}{\sqrt{d_k}} + E_{pos,i}\right) \tag{13}$$



MSCAS-Net
Multi-Scale Cross and Self-Attention Network
Fundus Image
Pre-Processing
12x12x1024
24x24x512
48x48x256
Swin Transformer
Conv1x1
Conv2x2
Conv4x4
Convolutional Feature Processing
12*12*1024
12*12*1024
12*12*1024
Self Attention
Self Attention
Self Attention
Cross Attention
Fuses features from self-attention blocks to capture global context.
Uses queries, keys, and values from self-attention outputs.
Multi-Scale Feature Extraction: Image is divided into patches at multiple scales:
Scale 1: 12x12x1024
Scale 2: 24x24x512
Scale 3: 48x48x256
Extracts features at each scale for richer information.
Applies convolutional layers (1x1, 2x2, 4x4) to adjust spatial dimensions.
Enhances feature representations while preserving scale consistency.
Self-Attention Mechanism Applied to each feature map to capture local dependencies
Classification
Unified Feature Vector
Combines local and global features into a single vector.
Represents the image in a compact form for classification in the form of 12x12x1024.

Fig. 1. MSCAS-Net (multi-scale cross and self-attention network) architecture.

.

Where $E_{pos,i}$ preserves positional information, and $\sqrt{d_k}$ stabilizes gradients during training. The refined feature representation for each level is then calculated as:

$$F_{\text{refined},i} = A_i \cdot V_i \tag{14}$$

.

This process allows the self-attention layers to emphasize relevant spatial features and suppress redundant information. Global features $F_{\text{refined},1}$ capture broad contextual patterns, mid-level features $F_{\text{refined},2}$ focus on intermediate structures, and local features $F_{\text{refined},3}$ highlight fine-grained details.

By leveraging self-attention, the model captures long-range dependencies across spatial locations, enhances important features, and integrates hierarchical positional information. This mechanism ensures robust feature refinement and improves the performance of downstream tasks by adaptively balancing global and local information as illustrated in Fig. 2.

Hierarchical feature fusion via cross-attention

After refining the multi-scale feature maps $F_{\text{refined},1}$, $F_{\text{refined},2}$ and $F_{\text{refined},3}$ Using self-attention, the next step is to fuse these hierarchical features into a single, unified representation. Cross-attention facilitates this fusion by dynamically aligning and integrating complementary information across different levels of abstraction. This mechanism's key point is to ensure that the fused feature map contains the global context, intermediate structures, and fine-grained details needed for good image classification.

To achieve this, the global feature map $F_{\text{refined},1}$ is utilized as the query, while the mid-level $F_{\text{refined},2}$ and local $F_{\text{refined},3}$ feature maps serve as the keys and values. Each feature map is projected into query Q, key K, and value V embeddings using learnable weight matrices:

$$Q = W_q \cdot F_{\text{refined},1}, i \ \{2{,}3\} \tag{15}$$

$$K_i = W_{k,i} \cdot F_{\text{refined},i}, \tag{16}$$

$$Vi = W_{v,i} \cdot F_{\text{refined},i}, \tag{17}$$

.

The cross-attention mechanism calculates the relevance between the global features and lower-scale features using scaled dot-product attention:

$$A_i = \text{Softmax}\left(\frac{Q \cdot K_i^{\top}}{\sqrt{d_k}} + E_{\text{pos},i}\right) \tag{18}$$

.

This process allows the global features to selectively emphasize the most relevant regions within the mid-level and local feature maps. The refined contributions from each scale are then computed as:

$$\mathrm{F}_{\mathrm{fusion,i}} = \mathrm{A_i} \cdot \mathrm{V_i}, \quad i \in \{2{,}3\} \tag{19}$$

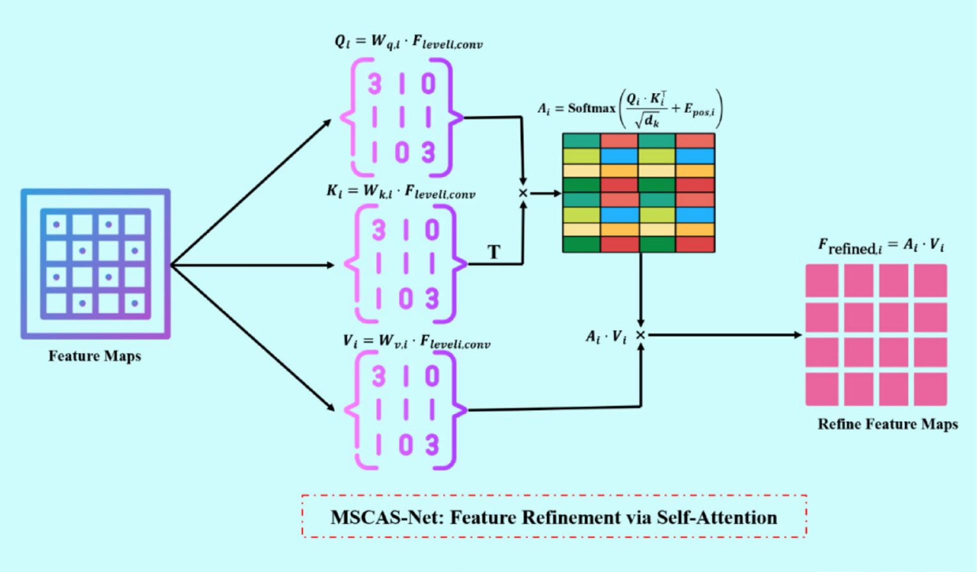

$Q_i = W_{q,i} \cdot F_{leveli,conv}$
3 1 0
1 1 1
1 0 3
$A_i = \text{Softmax}\left(\frac{Q_i \cdot K_i^T}{\sqrt{d_k}} + E_{pos,i}\right)$
$K_i = W_{k,i} \cdot F_{leveli,conv}$
×
3 1 0
1 1 1
1 0 3
T
$F_{refined,i} = A_i \cdot V_i$
Feature Maps
$A_i \cdot V_i$
×
$V_i = W_{v,i} \cdot F_{leveli,conv}$
3 1 0
1 1 1
1 0 3
Refine Feature Maps
MSCAS-Net: Feature Refinement via Self-Attention

Fig. 2. Self-attention block.

.

Finally, the fused representation is obtained by aggregating the global features and the refined outputs:

$$F_{\text{fused}} = F_{\text{refined},1} + F_{\text{fusion},2} + F_{\text{fusion},3} \tag{20}$$

.

This unified feature map balances global and local information, ensuring a comprehensive understanding of the input image. Cross attention augments the model's capability in capturing all the detail at a small scale without being concerned with the large pattern, with the help of adaptively aligned features across scale. Not only this, but this fusion strategy has the best classification accuracy and robustness for situations with different feature hierarchies as illustrated in Fig. 3.

Unified feature representation and multi-class classification

The hierarchical features are first refined with self-attention feature and combined with the cross-attention feature, then a Global Average Pooling (GAP) is used to simplify the multi-scale feature map F_{fused} to a unified feature vector F_{unified}. The spatial average of each feature channel is done by GAP, which effectively condenses the global and local contextual information to a compact and fixed-sized representation:

$$F_{\text{unified}}\left[c\right] = \frac{1}{H \cdot W}\sum\nolimits_{h=1}^{H}\sum\nolimits_{w=1}^{W} F_{\text{fused}}\left[c, h, w\right] \tag{21}$$

.

Where c signifies the channel index, H signify the height and W signify width of the feature map.

Then, the final class probabilities are obtained from the fully connected layer with softmax activation given the unified feature vector F_{unified}:

$$\widehat{y_i} = \text{Softmax}\left(W_{\text{cls}} \cdot\ F_{\text{unified}} + b_{\text{cls}}\right) \tag{22}$$

.

Thus, in the above, $\widehat{y_i}$ represents the predicted probability for class i corresponding to the learnable weights and biases of the classification layer W_{cls} and b_{cls}, respectively.

Then the categorical cross-entropy loss, a robust loss for multi class classification is used to optimize the model:

$$L_{\text{CCE}} = -\sum\nolimits_{i=1}^{C} y_i \log\left(\widehat{y_i}\right) \tag{23}$$

.

Where y_i is the ground truth label for the class i, $\widehat{y_i}$ is the predicted probability for the class i, and c is the total number of classes.

Where y_i is the ground truth label for the class i, $\widehat{y_i}$ is the predicted probability for the class i, and c is the total number of classes.

This unified approach guarantees that the model will be able to balance global and local information and elaborate hierarchical information while maintaining high discriminative power for many classes. GAP also

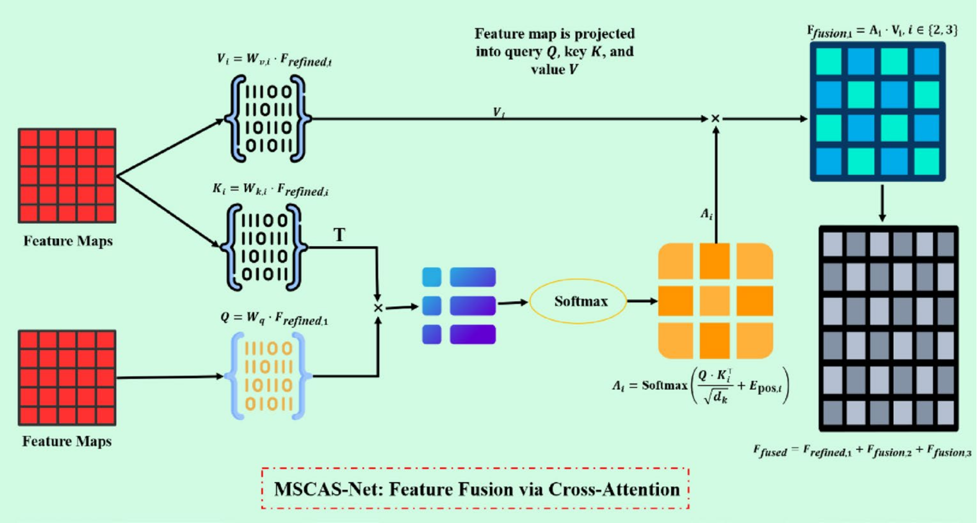

Feature map is projected into query Q, key K, and value V
$F_{fusion,i} = A_i \cdot V_i, i \in \{2,3\}$
$V_i = W_{v,i} \cdot F_{refined,i}$
V_i
Feature Maps
$K_i = W_{k,i} \cdot F_{refined,i}$
T
A_i
Softmax
$Q = W_q \cdot F_{refined,1}$
$A_i = \text{Softmax}\left(\frac{Q \cdot K_i^T}{\sqrt{d_k}} + E_{pos,i}\right)$
Feature Maps
$F_{fused} = F_{refined,1} + F_{fusion,2} + F_{fusion,3}$
MSCAS-Net: Feature Fusion via Cross-Attention

Fig. 3. Cross attention block.

contributes to reducing overfitting while maintaining spatial interpretability, which makes the model very suitable for complex tasks like medical imaging or fine-grained classification.

Experimental setup

Datasets

We assess the efficacy of our method using three publicly available datasets: APTOS, DDR, and IDRID. This subsection provides further information and details regarding these three datasets.

APTOS 2019 dataset[47]

The APTOS 2019 dataset, also known as the APTOS blindness detection dataset, was created by the Asia Pacific Tele-Ophthalmic Society (APTOS). It contains 3,662 retinal fundus images annotated according to the ICDR and the ETDRS scales. It separates the images into five severity levels reflecting No DR (categorised as 0), Mild (categorised as 1), Moderate (categorised as 2), Severe (categorised as 3), and Proliferative DR (categorised as 4). In the dataset, 1805 images belong to class 0, 370 images to class 1, 999 images to class 2, 193 images to class 3, 295 images to class 4. Image Resolutions of the images ranges from 640 × 480 to 2896 × 1944. The APTOS 2019 dataset was chosen for its diverse, real-world retinal fundus images with varying resolutions and quality, reflecting clinical scenarios. Its five DR severity levels and class imbalance, Image Noise and Inconsistent Illumination challenge model robustness, requiring robust preprocessing techniques like resizing, noise removal, and contrast enhancement. This clinically relevant benchmark validates MSCAS-Net's performance and supports the development of deep learning-based DR diagnosis systems.

IDRID[48]

It is a dataset based on the Indian diabetic retinopathy Image Dataset (IDRID), which contains images annotated for diabetic retinopathy severity at levels ranging from 0 (no retinopathy) to 4 (proliferative retinopathy). The data in this dataset can help develop and validate machine learning models and computer vision algorithms. This dataset IDRID is composed of 516 high-resolution retinal images in JPEG format and is used for training and evaluation. In the dataset, 134 images belong to class 0, 20 images to class 136 images to class 2, 74 images to class 3, 49 images to class 4. Image dimensions are 4288 × 2848, which provide high-quality visual data for detailed analysis. This small dataset is selected for its high-resolution and having detailed annotations for five DR severity levels. Its class imbalance and high-quality visuals test the model's ability to detect subtle lesions, ensuring precise fine-grained classification.

DDR[49]

The DDR Dataset is created for diabetic retinopathy classification and segmentation and serves as an essential tool for machine learning and deep learning research. High-resolution retinal fundus images are provided that include diabetic retinopathy severity levels ranging from 0 (healthy) to 4 (proliferative retinopathy). The dataset has about 12,522 images varying in representation of various stages of the disease. Among these images, 6266 images belong to class 0, 630 images to class 1, 4477 images to class 2, 236 images to class 3, 913 images to class 4. Image dimensions ranges from 512 × 512 to 5184 × 3456. The DDR dataset, annotated for DR classification and lesion segmentation, is selected for being a very large dataset with high variation in image quality and presence of noise, diverse image quality and class distribution. This versatility enables it to effectively evaluate MSCAS-Net's robustness and generalization, ensuring reliable performance across complex, real-world medical imaging challenges, making it an ideal benchmark for advanced diagnostic systems. In revolutionizing the diagnosis and treatment planning of human diseases, the DDR dataset provides a benchmark for the improvement of the automated diagnosis and treatment planning systems. Table 3 shows the data distribution for the purpose of Training, Validation, and Testing on three publicly available datasets.

Implementation details

In the experiments, we use the Swin Transformer for multiscale feature extraction for scales of 12 × 12, 24 × 24, 48 × 48 and generate feature maps. In each scale we do feature refinement with self-attention and try to fuse the multi scale features via cross attention. Given the enhanced feature map, the channel attention layer generates the attention maps. On the dataset, the number of training epoch is dependent, namely for DDR the number of training epoch was 40, IDRID was 40 and also APTOS was also 40. The batch size is set to 32. Initial learning rate is set to 0.0001, then an exponential decay with factor of 0.9 after specified epochs; and the weighting factors in Categorical cross entropy loss function is used to keep classification accuracy balanced. The experiments were carried out on an NVIDIA RTX 4060ti GPU using PyTorch. In the near future, the code will be available for download.

Purpose	APTOS	IDRID	DDR
Total images	3662	516	12,522
Training, validation	80% (2930)	80% (413)	80% (10018)
Testing	20% (732)	20% (103)	20% (2504)

Table 3. Data distribution for training, validation, and testing across APTOS, IDRID, and DDR datasets.

Research questions

RQ1:How can advanced feature extraction methods address the complexity of lesions in fundus images from the APTOS 2019 and DDR datasets?

RQ2: How can models be effectively trained on the APTOS 2019 and IDRID datasets despite class imbalance?

RQ3: What augmentation techniques can improve model generalization on the APTOS 2019 and IDRID datasets?

RQ4: How can models be made robust to the diverse dimensions and quality of images in the APTOS 2019 and DDR datasets?

Challenges in feature extraction

Punctual and circumscribed lesions in the APTOS 2019 and DDR datasets are somewhat difficult to extract from fundus images. To this aim, the MSCAS-Net model proposed in this paper uses the multi-scale feature extraction where images are divided into patches of 1 × 1, 2 × 2, 4 × 4. This enables the model to learn the subtle details of the lesions as well as broad features of the images. The spatial variant of attention applied to each feature map is dedicated to capturing local dependencies. In contrast, the cross-variant of attention focuses on global context by using queries, keys, and values derived from the self-attention outputs. This enables the model to obtain different scale features and strengthen their interaction, while testing on the DR image data leads to consideration of multiplex lesion features.

Dataset imbalance

The datasets of APTOS 2019 and IDRID have imbalanced classes; that is, during training, the model may be inclined to predict the majority class rather than the actual incidence since it is more frequent. Through the multi-scale features extraction coupled with attention mechanism, MSCAS-Net addresses this problem by improving the net's capacity to learn from minority classes. In this way, extracting features at different scales may help the model to concentrate on tiny lesions that are generally dissimilar and not very distinguishable in imbalanced datasets. On the same note, the cross-attention helps to keep a track of the global context and hence, does not easily overspecialize for the most frequent classes. It is also worth emphasizing that the proposed model takes both local and global features into account and combines them into the cultivable feature vector, thus, enhancing classification accuracy in all the severity levels of DR.

Need for dataset augmentation

Augmentation is needed for the APTOS 2019 and IDRID datasets so that model generalization can be improved. An inherent robustness to dataset variations is introduced into MSCAS-Net by the multi-scale feature extraction. The model learns invariant features that are less sensitive to image orientation, size, and quality variations by processing images at multiple scales (12 × 12, 3 × 4, 4 × 4). It also enables the model to concentrate on the good things in the data no matter how small it is. Somehow, the model can generalize from the data at hand, as its multi-scale and attention-based architecture decreases the reliance on the heavy augmentation techniques.

Variability in image quality

The APTOS 2019 and DDR datasets consist of images of varying sizes and image quality, which makes it difficult to reach the optimum performance of the model. To address this issue, MSCAS-Net features a multi-scale feature extraction and 1 × 1, 2 × 2, 4 × 4 layer adjustment in its convolutional layers. The model processes the images in multiple scales, allowing it to work with varying resolutions and quality. In addition, the self-attention mechanism further helps us concentrate on correct features in low-quality images. The cross-attention mechanism also adds the global context compensation to deal with the disparity in image quality characteristics. Therefore, providing robust performance at both multi-scale and attention scales, the model shows a high level of effectiveness for DR classification.

Ablation study

Effect of Swin transformer as backbone

MSCAS-Net implements Swin Transformer as its core component because this architecture utilizes self-attention models to extract both local and global image characteristics needed to diagnose DR conditions. The network divides images into multi-scale patches (12 × 12, 24 × 24, 48 × 48) to extract different-level features needed for detecting delicate lesions. The framework maintains simultaneous visibility of both small features (48 × 48) and sizeable contextual information. The model utilizes shifted window-based self-attention to detect dependencies between distant elements, which is essential for medical image analysis, particularly when studying lesions with different size placements. High-resolution fundus image analysis becomes possible with the Swin Transformer because of its efficient computation. MSCAS-Net integration delivers exceptional performance results on both APTOS and IDRID database examinations.

Effect of self-attention for multiscale feature enhancement

Self-attention operates on each scale independently to define spatial patterns for detecting discriminative elements (for example, hemorrhages) and minimizing noise in the process. Within the 12 × 12 pixel grid, Self-attention analyzes broader context patterns related to edema, and in the 48 × 48 scale, it detects weak lesions in the image. Primary texture similarity and background interference problems are solved by the hierarchical refinement method that yields robust features in variations in resolution.

Effect of cross-attention for feature fusion

Through cross-attention, the system merges details from various feature scales by connecting small-scale lesion boundary detection with large-scale disease severity assessment. By using global features as queries, together with mid/local features as keys/values, the model detects relationships between different scales where faint lesions find connections to broader pathological contexts. The integration of different data levels minimizes unbalanced scenarios in the dataset and eliminates useless data by identifying only relevant medical information. Because of this integration, the system obtains superior resistance against changes in image quality while simultaneously achieving higher diagnosis precision.

The Table 4 presents the accuracy achieved by the Swin Transformer model with varying attention mechanisms—Self Attention and Cross Attention—across three datasets: APTOS, IDRID, and DDR. The results indicate that the combination of both attention types yields the highest classification accuracy for each dataset.

Evaluation metric

Accuracy

Accuracy measures the proportion of correct predictions (both true positives and true negatives) to the total predictions. It is useful when the dataset is balanced.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{FP} + \boldsymbol{TN} + \boldsymbol{FN}} \tag{24}$$

.

Where:

TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

Experimental results

Thus, this section regards the explanation, evaluation, and analysis of our newly proposed approach via a couple of metrics on four datasets (APTOS, DDR, IDRID). Moreover, the usefulness of the pre-processing step towards improving the achieved accuracy is shown. Additionally, our results are compared with the current state-of-the-art approaches for label grading based on DR severity. We applied our approach using the TensorFlow framework, with the Adam optimizer used to train the model with a learning rate of 1e-4. Moreover, we set the batch size to be 16 in order to train our model shown in Figs. 4, 5, 6, 7, 8 and 9, and 10; Table 4.

Table 5 compares the proposed MSCAS model against state-of-the-art (SOTA) methods for DR grading across three datasets: APTOS 2019, DDR, and IDRID. The results demonstrate that the MSCAS model achieves impressive accuracy rates of 93.8% for APTOS 2019, 89.80% for DDR, and 86.70% for IDRID. These findings highlight the effectiveness of the MSCAS model in outperforming several established studies, such as those by Mondal et al. and Bodapati et al. The significant improvements in accuracy underscore the model's potential contributions to enhancing DR classification strategies and its applicability in clinical settings.

Discussion

MSCAS-Net shows exceptional performance for the DR classification because of its effective operations. This model integrates multilevel feature extraction methods combined with advanced attention techniques to achieve better results, that is, retaining the details within its local information while processing global elements. MSCAS-Net delivers top performance across different benchmark datasets, which include APTOS, DDR, and IDRID. The model achieves exceptional outcomes when it deals with unbalanced datasets and variable image quality. The system faces multiple hurdles even after its advancements have been made. The feature weight balance across different scales extracted through self-attention and cross-attention mechanisms needs better optimization in the model construction process. The model has lower dependence on data augmentation yet requires additional improvements to maintain robustness, particularly for real-world image issues which present extreme quality problems. The clinical use of this model depends heavily on its clear interpretation capabilities. Additional research should concentrate on improving model understanding for better assistance of medical choices. The model has a computational cost of 9.2 GFLOPs and contains ~ 31 million parameters when processing standard 224 × 224 input images. These metrics ensure efficient real-time performance on modern GPUs, making MSCAS-Net suitable for clinical deployment. MSCAS-Net represents a dependable automated DR diagnostic solution while establishing groundwork for study regarding multi-scale features and attention mechanism development.

Swin transformer	Self attention	Cross attention	Accuracy achieved		
			APTOS (%)	IDRID (%)	DDR (%)
✓	–	–	75.50	69.20	71.30
✓	✓	–	86.20	81.01	83.66
✓	–	✓	89.52	84.6	87.53
✓	✓	✓	93.80	86.70	89.80

Table 4. Accuracy achieved by different attention mechanisms in the ablation study for various datasets (APTOS, IDRID, DDR).

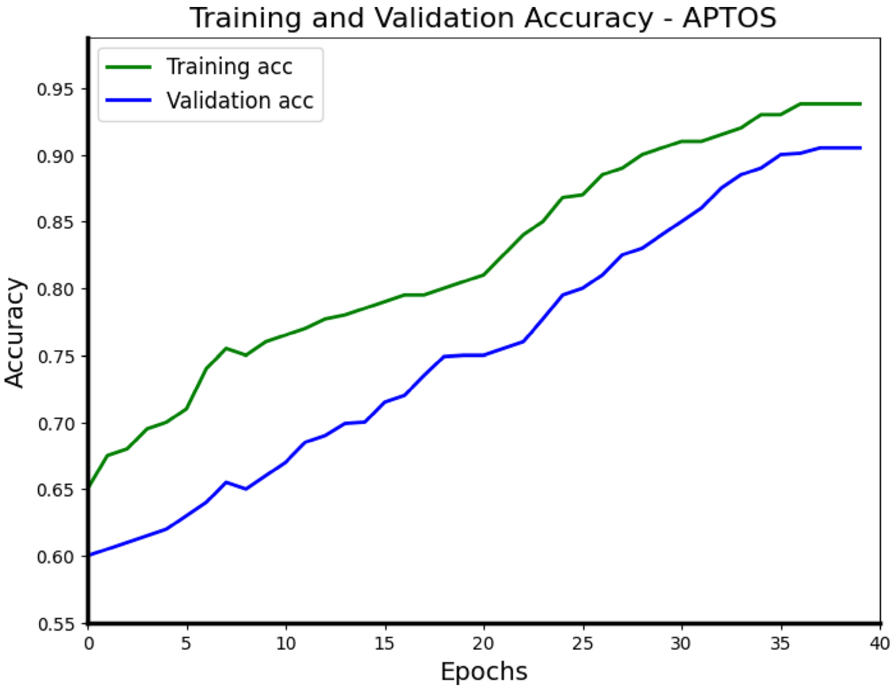


Fig. 4. Training and validation accuracy curves for APTOS dataset.

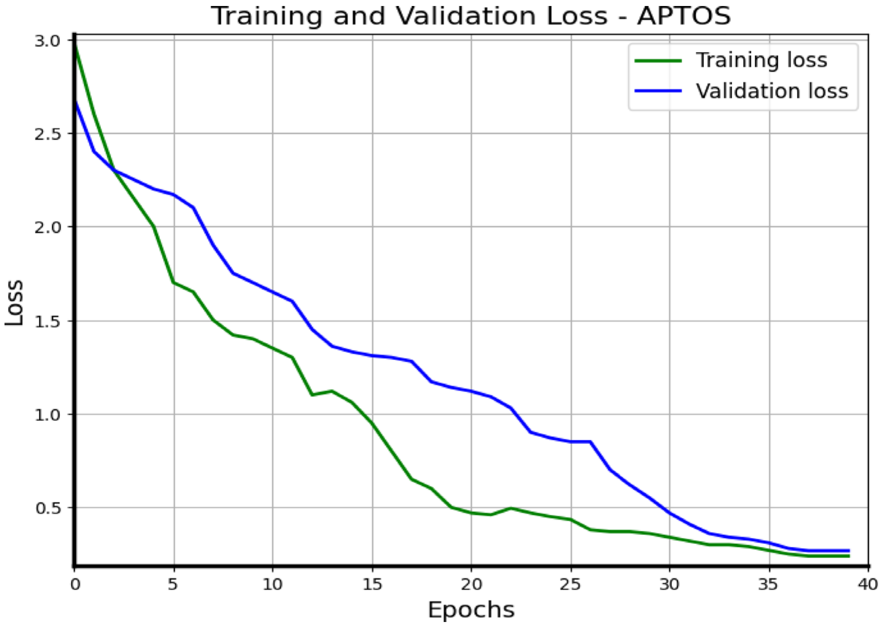


Fig. 5. Training and validation loss curves for APTOS dataset.

Conclusion

We present the Multi-Scale Cross and Self-Attention Network (MSCAS-Net) in this work for the fine-grained diabetic retinopathy (DR) classification, which, being a challenging problem in fine grained visual classification, leads to mitigate important issues. MSCAS-Net enhances its capability to recognize subtle lesions of high importance for accurate diagnosis by capturing local and global features of retinal images using advanced multi scale feature extraction and attention mechanisms. Extensive experiments on the APTOS, DDR, and IDRID datasets show that MSCAS-Net achieves the best performance in classification accuracy and leads to good per-class performance with effectiveness in both dataset imbalance and image variability problems. By integrating complementary information across the scales, through the combination of self-attention and cross-attention mechanisms in the model architecture, it is possible to focus on the relevant features and achieve a

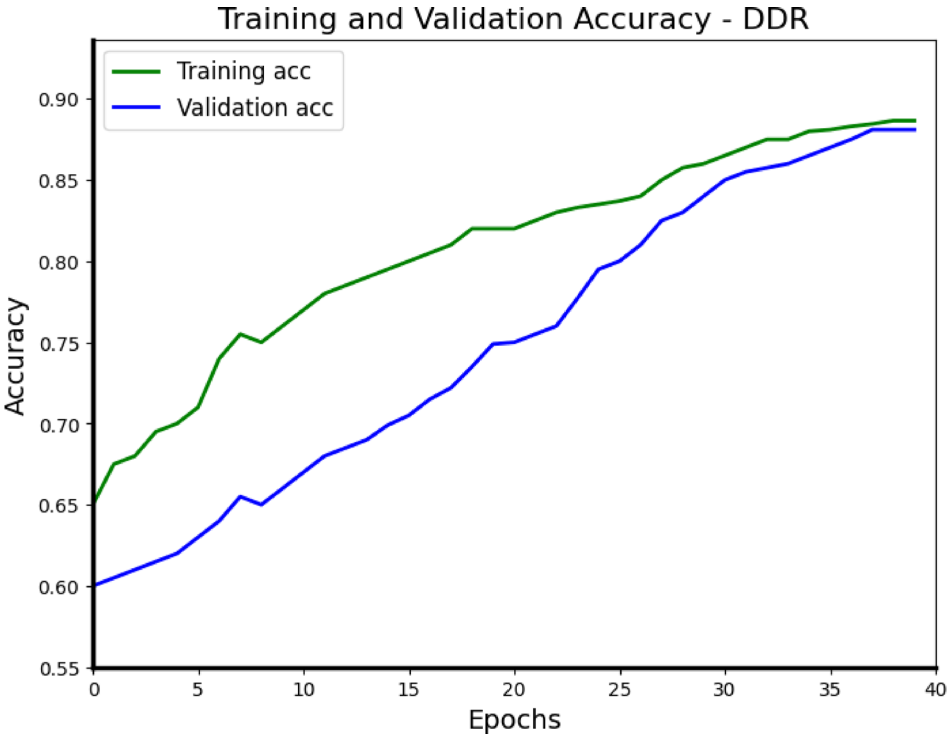


Fig. 6. Training and validation accuracy curves for DDR dataset.

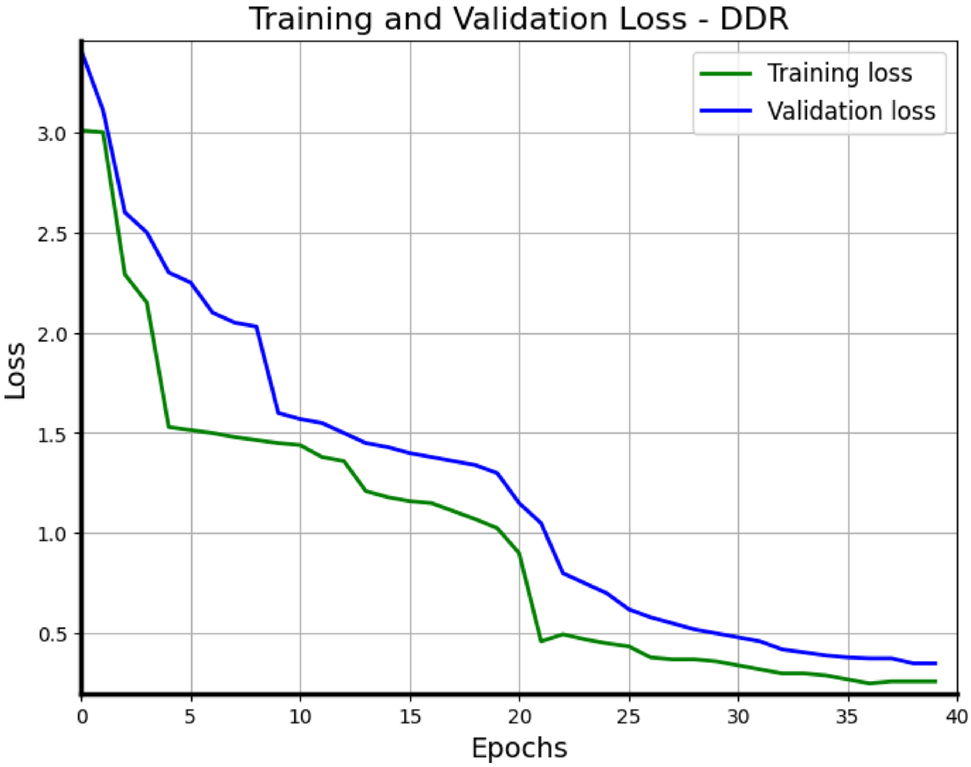


Fig. 7. Training and validation loss curves for DDR dataset.

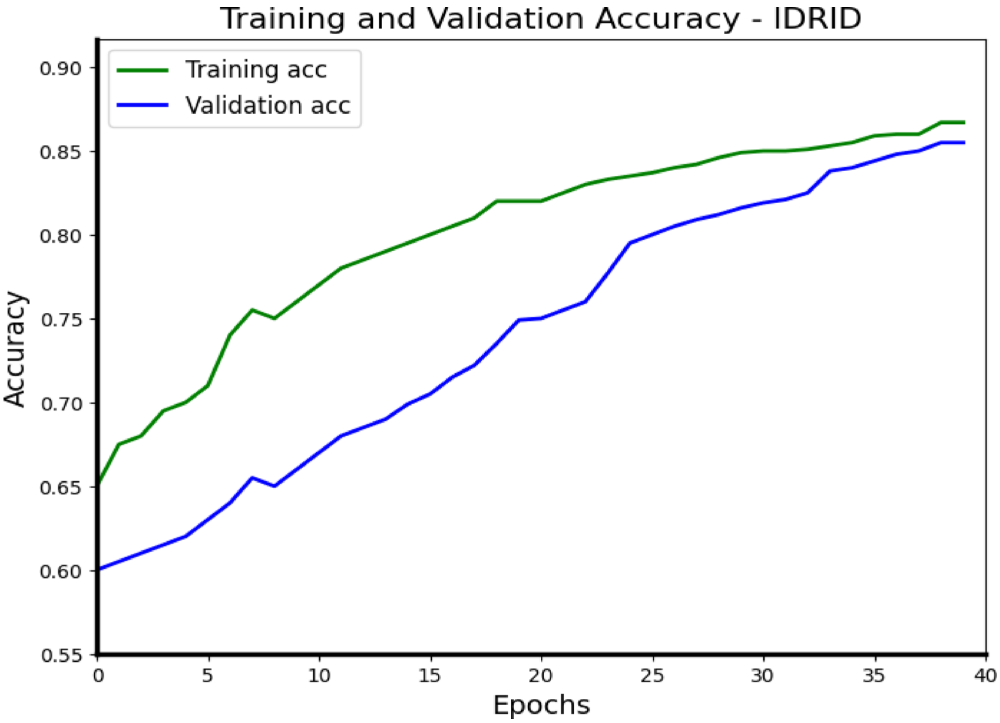


Fig. 8. Training and validation accuracy curves for IDRID dataset.

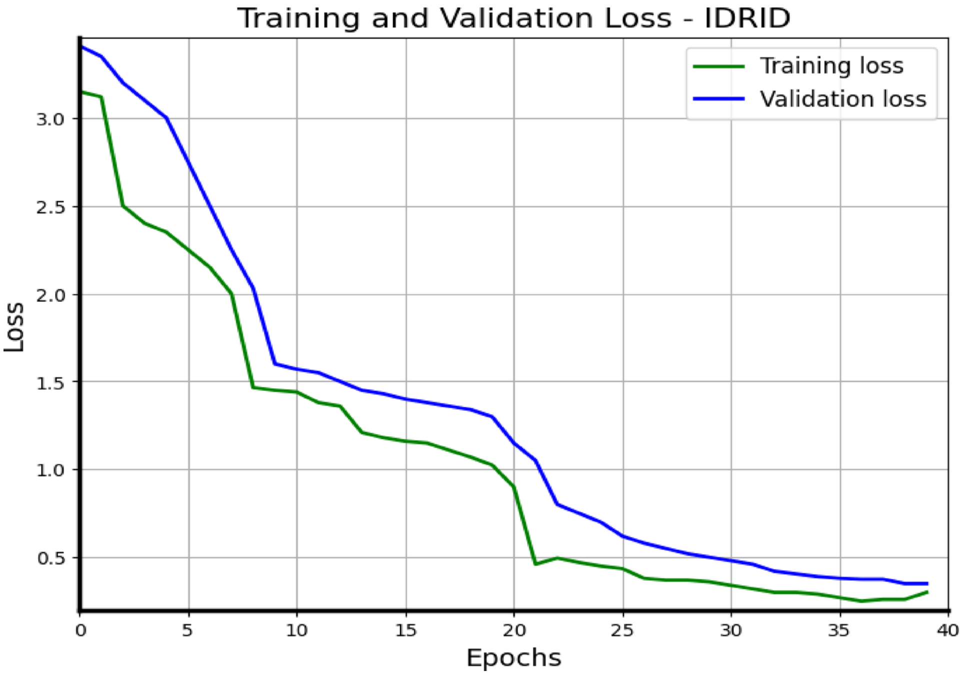


Fig. 9. Training and validation loss curves for IDRID dataset.

robust feature representation. Also, the multi-scale approach is less sensitive to image orientation and quality, thus, the performance is less sensitive to variation in image orientation and quality for different datasets. The potential application of the model in real-world clinical settings is one burden of future work aimed at increasing model interpretability. The promising results of this research provide the basis for using advanced deep learning techniques in medical imaging to enhance image-based automated diagnostic systems for diabetes retinopathy and other ophthalmological diseases. However, MSCAS-Net stands as a significant leap from existing work of accurate and efficient trustworthy AI healthcare solutions.

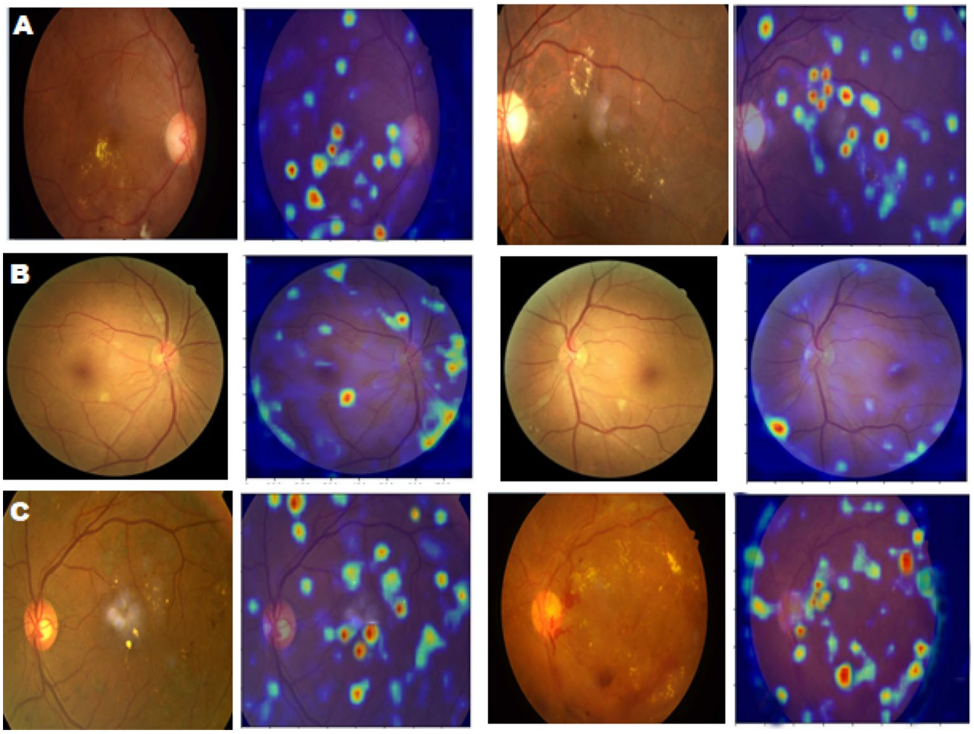


Fig. 10. A visual example of heatmaps generated using this model on three datasets (**a**) APTOS (**b**) DDR and (**c**) IDRID.

Dataset	Study	DR grading (classes)	Accuracy %
APTOS 2019	Mondal et al.[7]	5	86.08
	Vijayan et al.[15]		86.20
	Bodapati et al.[35]		84.17
	Shaik and Cherukuri[38]		85.54
	Proposed		93.8
DDR	Zhao et al.[9]	5	83.10
	Vijayan et al.[15]		84.80
	Mubashra[45]		89.29
	Oulhadj[46]		80.36
	Proposed		89.80
IDRID	Bodapati et al.[35]	5	63.24
	Shaik and Cherukuri[38]		66.41
	Jiwani[50]		77.60
	Santos et al.[51]		77.50
	Proposed		86.70

Table 5. Comparison with SOTA methods with proposed MSCAS model.

Data availability

Availability data and MaterialThis paper contains all the data in the links given below.APTOShttps://www.kaggle.com/datasets/mariaherrerot/aptos2019IDRIDhttps://www.kaggle.com/datasets/mariaherrerot/idrid-dataset DDR• Dataset Link: https://www.kaggle.com/datasets/mariaherrerot/ddrdataset.

Received: 14 March 2025; Accepted: 16 June 2025
Published online: 01 July 2025

References

1. Salud, O. M. d.l. *Organización Mundial de la Salud*. https://www.who.int/es/news-room/fact-sheets/detail/diabetes.

2. Hegde, A. & Sumana, K. R. Comparative study of diabetic retinopathy detection using machine learning techniques. *Int. J. Res. Appl. Sci. Eng. Technol.* (2022).
3. Wan, S., Liang, Y. & Zhang, Y. Deep convolutional neural networks for diabetic retinopathy detection by image classification. *Computers Electr. Eng.* **72**, 274–282 (2018).
4. Abbas, Q. et al. HDR-EfficientNet: a classification of hypertensive and diabetic retinopathy using optimize Efficientnet architecture. *Diagnostics* **13**(20), 3236 (2023).
5. Harithalakshmi, K., Rajan, R. & Nadheera, K. EfficientNet-based diabetic retinopathy classification using data augmentation. In *2023 9th International Conference on Smart Computing and Communications (ICSCC)*. (IEEE, 2023).
6. Li, X. et al. CANet: cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading. *IEEE Trans. Med. Imaging*. **39**(5), 1483–1493 (2019).
7. Mondal, S. S. et al. Edldr: an ensemble deep learning technique for detection and classification of diabetic retinopathy. *Diagnostics* **13**(1), 124 (2022).
8. Wang, Z. et al. Generative adversarial networks in ophthalmology: what are these and how can they be used? *Curr. Opin. Ophthalmol.* **32**(5), 459–467 (2021).
9. Zhao, S. et al. CoT-XNet: contextual transformer with Xception network for diabetic retinopathy grading. *Phys. Med. Biology*. **67**(24), 245003 (2022).
10. Dihin, R. A., Al-Jawher, W. A. M. & AlShemmary, E. N. Diabetic retinopathy image classification using shift window transformer. *Int. J. Innovative Comput.* **13**(1–2), 23–29 (2022).
11. Madarapu, S., Ari, S. & Mahapatra, K. A deep integrative approach for diabetic retinopathy classification with synergistic channel-spatial and self-attention mechanism. *Expert Syst. Appl.* **249**, 123523 (2024).
12. Lin, C. L. & Wu, K. C. Development of revised ResNet-50 for diabetic retinopathy detection. *BMC Bioinform.* **24**(1), 157 (2023).
13. Asia, A. O. et al. Detection of diabetic retinopathy in retinal fundus images using CNN classification models. *Electronics* **11**(17), 2740 (2022).
14. Boruah, S. et al. Gaussian blur masked resnet2. 0 architecture for diabetic retinopathy detection. *Computers Mater. Continua*. **75**(1), 927–942 (2023).
15. Vijayan, M. A regression-based approach to diabetic retinopathy diagnosis using Efficientnet. *Diagnostics* **13**(4), 774 (2023).
16. Islam, M. R. et al. Applying supervised contrastive learning for the detection of diabetic retinopathy and its severity levels from fundus images. *Computers Biology Med.* **146**, 105602 (2022).
17. Alharbey, R. et al. Indexing important drugs from medical literature. *Scientometrics* **127**(5), 2661–2681 (2022).
18. Abbas, T. et al. IoMT-based healthcare systems: A review. *Comput. Syst. Sci. Eng.*, **48**(4) (2024).
19. Yang, Y. et al. A novel transformer model with multiple instance learning for diabetic retinopathy classification. *IEEE Access*. **12**, 6768–6776 (2024).
20. Masood, I. et al. A blockchain-based system for patient data privacy and security. *Multimedia Tools Appl.* **83**(21), 60443–60467 (2024).
21. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2021).
22. Badar, D. et al. Automated dual CNN-based feature extraction with SMOTE for imbalanced diabetic retinopathy classification. *Image Vis. Comput.* 105537 https://doi.org/10.1016/j.imavis.2025.105537 (2025).
23. Hayat, M. K. et al. Towards deep learning prospects: insights for social media analytics. *IEEE Access*. **7**, 36958–36979 (2019).
24. Han, K. et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(1), 87–110 (2022).
25. Masood, I. et al. Towards smart healthcare: patient data privacy and security in sensor-cloud infrastructure. *Wirel. Commun. Mob. Comput.* **2018**(1), 2143897 (2018).
26. Wei, X. S. et al. Fine-grained image analysis with deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(12), 8927–8948 (2021).
27. Yan, Y. AlexViT: novel diabetic retinopathy image classification. In *2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI)*. (IEEE, 2023).
28. Li, W. et al. Interpretable detection of diabetic retinopathy, retinal vein occlusion, Age-Related macular degeneration, and other fundus conditions. *Diagnostics* **14**(2), 121 (2024).
29. Javed, R. et al. Deep learning for lungs cancer detection: a review. *Artif. Intell. Rev.* **57**(8), 197 (2024).
30. Huang, Z. & Li, Y. Interpretable and accurate fine-grained recognition via region grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2020).
31. Zhou, M. et al. Look-into-object: Self-supervised structure modeling for object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2020).
32. Zhu, H. et al. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2022).
33. Oulhadj, M. et al. Diabetic retinopathy prediction based on deep learning and deformable registration. *Multimedia Tools Appl.* **81**(20), 28709–28727 (2022).
34. Oulhadj, M. et al. Diabetic retinopathy prediction based on wavelet decomposition and modified capsule network. *J. Digit. Imaging*. **36**(4), 1739–1751 (2023).
35. Bodapati, J. D. Stacked convolutional auto-encoder representations with spatial attention for efficient diabetic retinopathy diagnosis. *Multimedia Tools Appl.* **81**(22), 32033–32056 (2022).
36. Fan, R., Liu, Y. & Zhang, R. Multi-scale feature fusion with adaptive weighting for diabetic retinopathy severity classification. *Electronics* **10**(12), 1369 (2021).
37. Sugeno, A. et al. Simple methods for the lesion detection and severity grading of diabetic retinopathy by image processing and transfer learning. *Computers Biology Med.* **137**, 104795 (2021).
38. Shaik, N. S. & Cherukuri, T. K. Hinge attention network: A joint model for diabetic retinopathy severity grading. *Appl. Intell.* **52**(13), 15105–15121 (2022).
39. Al-Antary, M. T. & Arafa, Y. Multi-scale attention network for diabetic retinopathy classification. *IEEE Access*. **9**, 54190–54200 (2021).
40. Abbasi, S. et al. Classification of diabetic retinopathy using unlabeled data and knowledge distillation. *Artif. Intell. Med.* **121**, 102176 (2021).
41. Hu, Y. et al. Rams-trans: Recurrent attention multi-scale transformer for fine-grained image recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*. (2021).
42. Demidov, D. et al. *Salient mask-guided vision transformer for fine-grained classification*. arXiv preprint arXiv:.07102, (2023).
43. Gao, Y. et al. Channel interaction networks for fine-grained image categorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*. (2020).
44. Rao, Y. et al. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2021).
45. Mubashra, A., Naeem, A., Aslam, D. N., Abid, M. K. & Haider, J. Diabetic retinopathy identification from eye fundus images using deep features. *VFAST Trans. Softw. Eng.* **11**(2), 172–186 https://doi.org/10.21015/vtse.v11i2.1206 (2023).
46. Oulhadj, M. et al. Diabetic retinopathy prediction based on vision transformer and modified capsule network. *Comput. Biol. Med.* **175**, 108523. https://doi.org/10.1016/j.compbiomed.2024.108523 (2024).

47. Karthik, M. S. D. *APTOS 2019 Kaggle Diabetic Retinopathy Detection Competition Dataset*. https://www.kaggle.com/c/aptos2019-blindnessdetection/data (2019).
48. Porwal, P. et al. Idrid: Diabetic retinopathy–segmentation and grading challenge. **59**, 101561. (2020).
49. Li, T. et al. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. **501**, 511–522. (2019).
50. Jiwani, N. et al. Application of transfer learning approach for diabetic retinopathy classification. In *Proceedings – 2nd International Conference on Power Electronics and Energy, ICPEE 2023*,https://doi.org/10.1109/ICPEE54198.2023.10060777 (2023).
51. Santos, M. S., Valadao, C. T., Resende, C. Z. & Cavalieri, D. C. Predicting diabetic retinopathy stage using Siamese convolutional neural network. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **12**(1). https://doi.org/10.1080/21681163.2023.2297017 (2024).

Author contributions

D.B., T.A., J.A., and R.A. collected data from different resources. T. A., R. A., D. B., and J.A. contributed to writing—original draft preparation; A.D., T.A., R.A., and D. B. contributed to writing—review and editing. W. C., A.D. and R.A. supervised the paper. J.A., D.B., and R.A. drafted pictures and tables. D.B., and A.D. performed revisions and improved the quality of the draft. All authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.C. or A.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.