

Attention-Guided InceptionNeXt–Transformer Framework for Automated Diabetic Retinopathy Grading and Lesion Localization

Dr. T. Ilam Parithi¹, P. Nithya^{2*}

¹Guest Lecturer, Government Arts and Science College, Sankarankovil, Tenkasi, Tamil Nadu

²Research Scholar, Dr. N.G.P. Arts and Science College, Coimbatore

*Corresponding Author: P. Nithya, Research Scholar, Dr. N.G.P. Arts and Science College, Coimbatore

ABSTRACT

Diabetic retinopathy (DR) is a leading retinal disease that may cause progressive visual loss and blindness if not detected and treated early. Automated analysis of retinal fundus images can help in consistent assessment of disease severity and identification of clinically relevant retinal lesions. In this paper, an Attention-Guided InceptionNeXt-Transformer Framework for Automated Diabetic Retinopathy Grading and Lesion Localization using Indian Diabetic Retinopathy Image Dataset (IDRiD) is proposed. In the proposed framework, we combine the ability of Transformer-based learning to model global contextual relationships with the multi-scale feature extraction capability of InceptionNeXt and design attention mechanisms to emphasize discriminative retinal regions. To perform comparative analysis, several deep learning architectures were considered and a hybrid InceptionNeXt-Transformer framework is implemented for diabetic retinopathy grading. A YOLO based detection component is used for lesion analysis to localize clinically relevant retinal abnormalities. In addition, we integrate the visualization based on Grad-CAM to provide interpretable evidence of the regions affecting the model prediction. The main goal of the framework is to assure the integration of the grading of retinal diseases, the localization of lesions and the explainable automated analysis, allowing to assess the diabetic retinopathy through computer-aided approaches.

Keywords: Diabetic Retinopathy, IDRiD, Retinal Fundus Imaging, InceptionNeXt, Vision Transformer, Attention Mechanism, Lesion Localization, Deep Learning, Grad-CAM, Explainable Artificial Intelligence.

How to cite this article: Ilam Parithi T, Nithya P., Attention-Guided InceptionNeXt–Transformer Framework for Automated Diabetic Retinopathy Grading and Lesion Localization. Int J Drug Deliv Technol. 2026;16(86s): 2089-2100; DOI: 10.25258/ijddt.16.86s.189

Source of support: Nil.

Conflict of interest: None

1. Introduction

Diabetic retinopathy (DR) is a progressive retinal complication of diabetes mellitus and one of the main causes of preventable visual impairment. The clinical evaluation of DR has traditionally depended on detecting and grading retinal abnormalities seen on fundus photographs. The development of standardized clinical severity scales was fundamental to the establishment of a uniform framework for the assessment of DR, which provided a definition of disease severity based on the distribution of characteristic retinal lesions. A clinical scale developed by Wilkinson et al. is internationally accepted to grade diabetic macular edema and DR. This scale provides an organized framework for assessing the severity of the disease and makes for a consistent clinical interpretation [1]. Previous stages of computer-aided retinal analysis were mainly focused on the recognition of pathological abnormalities and retinal structures using image-processing

techniques. For example, the automatic analysis of retinal vessels was an important basis for the future development of computer-assisted ophthalmic diagnosis. Staal et al. presented a ridge-based method for retinal vessel segmentation in colour fundus images and showed that it is possible to computationally extract clinically relevant vascular structures [8]. Automated detection of small retinal abnormalities, in particular microaneurysms which are one of the earliest visible signs of DR, became one of the major research directions. Niemeijer et al. performed a study on automated microaneurysm detection through Retinopathy Online Challenge to set a crucial benchmark for computer-aided retinal lesion analysis [7]. With the increasing availability of retinal image databases and the progress of machine learning, the field gradually shifted from handcrafted image processing to learned feature representations. In the past, retinal computer-aided systems heavily relied on manually designed features and conventional classifiers. Learning-based approaches have begun to reduce the

dependence on explicit feature engineering. Subsequent research explored machine-learning frameworks capable of recognizing multiple pathological structures integrated lesion-aware frame that can recognize multiple pathological structures. Sil Kar and Maity showed the potential of computational lesion analysis in the assessment of retinal images by developing an automated method for the detection of retinal lesions for DR screening [9]. In a following work on deep learning for retinal microaneurysm detection, Dai et al. further validated the usefulness of learned representations for the detection of small retinal lesions by using clinical-report-guided information and multi-sieving feature learning [10]. The use of convolutional neural networks (CNNs) for retinal image analysis represented a significant shift. CNNs allow for the direct learning of hierarchical features from fundus photographs, and the substantial reduction of the need for manually engineered descriptors. Deep learning based retinal screening showed the ability to automatically detect diabetic retinopathy from colour fundus images. Quellec et al. investigated the application of deep feature learning in large scale retinal image analysis and deep image mining for DR screening [5]. Since then, Gargeya and Leng have also shown automated DR identification using deep learning, thus demonstrating the potential of learned image representations for automated retinal screening [2]. Autonomous artificial-intelligence-based diagnostic systems have provided further evidence of the clinical importance of automated retinal screening. For instance, Abramoff et al. investigated the detection of DR in primary-care environment [3].

Despite these advances, automated analysis of DR is still challenging due to the large variability in size, shape, colour, contrast and spatial distribution of retinal abnormalities. Microaneurysms, as small lesions, may occupy only a very limited portion of a fundus image whereas haemorrhages and exudates, as larger abnormalities, may have a significant morphological variation. Moreover, the DR grading process is more complicated than binary disease detection because it requires the differentiation of different disease severity stages. Evaluation of automated DR systems is further complicated by the difficulty to establish reliable reference standards and the variability among human graders [4]. Thus, to enable a reliable automated analysis, we need models that can extract fine-grained lesion information and at the same time represent the broader retinal context. The development of specific retinal datasets has provided a crucial basis for tackling these challenges. The Indian Diabetic Retinopathy Image Dataset (IDRiD) is a dataset created for the purpose of research in automated DR screening and retinal image analysis. The dataset includes expert

annotations of DR severity and clinically relevant retinal lesions, allowing both image-level grading and lesion-level analysis [6]. The related IDRiD challenge [6] had three main tasks: localization and segmentation of retinal landmarks, DR severity grading and retinal-lesion segmentation. The multi-task nature of IDRiD makes it fit well with the frameworks that attempt to bridge the gap between disease-level classification and lesion-level localization instead of viewing DR grading only as an isolated classification task. The advent of attention mechanisms has also opened a new important direction in retinal image analysis. Attention mechanisms can selectively focus on the informative features at the spatial or channel level. Traditional convolutional operations mainly collect information through local receptive fields. Therefore, attention-based retinal frameworks have been explored to improve the representation of subtle pathological features and to address class imbalance and fine-grained severity differences [28]. Attention-guided approaches are particularly relevant for DR as clinically important abnormalities can be present in small regions in a relatively large retinal background. Transformer-based architectures have recently improved the ability of deep learning models to encode long-range relations in images. Vision Transformer approaches demonstrate that image patches can be processed by self-attention mechanisms to model the global contextual relationships. Later, transformer-based methods have been applied to DR grading with models aiming to predict the severity of DR using fundus photographs [25]. More recent approaches combine transformer-based contextual modelling and convolutional feature extraction, which suggests that transformers and CNNs are complementary. Later research has investigated the hybrid architectures and enhanced attention mechanisms for a more detailed assessment of DR severity [24], [26]. Yang et al. proposed a transformer-based method for DR classification. Consequently, the trend of DR research is shifting from conventional CNN-only classification to architectures with features such as multi-scale feature extraction, attention, global contextual modelling, lesion-level analysis, and explainability. Recently, hybrid CNN–Transformer architectures have been investigated for attention-based feature fusion and DR severity assessment [38], [39]. Attention mechanisms, multi-scale feature fusion, CNN–Transformer integration, lesion-level analysis, and explainable artificial intelligence are also identified as important directions in automated DR analysis in recent reviews [17], [18]. This evolution from classical image processing and handcrafted features to CNNs, Vision Transformers, explainable AI, and lesion-level segmentation is also summarized in a recent comprehensive survey that reports on research from

the early stages of automated retinal analysis up to recent deep learning methods [20].

As automated systems have evolved toward clinically oriented retinal analysis, the importance of explainability has grown. High classification accuracy alone is not enough to say that the model has decided based on medically relevant retinal structures. Grad-CAM and other visualization methods can offer spatial evidence of areas that contribute to a prediction, which can help to assess if a model addresses relevant retinal abnormalities [31]. The grading component gives an estimate of the severity of DR while lesion localization identifies spatial regions associated with pathological changes. Combining classification and lesion localization can provide complementary information. The IDRiD dataset is especially suitable for this integrated perspective because its annotations allow for both disease grading and lesion-level analysis [6]. Thus, the proposed Attention-Guided InceptionNeXt-Transformer Framework for Automated Diabetic Retinopathy Grading and Lesion Localization integrates the Transformer-based contextual modelling with the multi-scale convolutional representation learning. The transformer module is able to capture more global contextual relationships in retinal images and InceptionNeXt provides an efficient multi-scale convolutional representation. Attention-guided feature processing is implemented to put emphasis on diagnostically informative regions. The framework also considers lesion localization to improve the interpretability of image-level DR grading and incorporates Grad-CAM-based visualization. Thus, the overall approach is a unified framework for retinal image analysis that combines DR grading, retinal lesion localization, attention-guided feature learning, hybrid CNN-Transformer representation, and explainable prediction.

This framework is intended not to substitute clinical assessment but to offer a comprehensive computational method to aid in automated screening and analysis of the retina. The field has evolved from standardized clinical grading [1] to automated retinal-structure analysis [8], lesion detection [7], deep learning-based DR recognition [2], specialized multi-task datasets such as IDRiD [6], and the latest CNN-Transformer and attention-based approaches [38], [39], indicating a clear trend toward more holistic retinal analysis systems. This progression is then followed by the proposed framework by integrating disease severity assessment, lesion localization and interpretable visual analysis.

II Literature Review

The research on automated diabetic retinopathy (DR) analysis has evolved from traditional retinal image-processing methods to deep learning, attention mechanisms, Transformer architectures and integrated lesion-aware frameworks. The first computer-aided analysis of retinal images had the primary objective of the extraction of anatomical structures and pathological abnormalities from fundus pictures. Segmentation of retinal vessels was an important foundation, since vascular structures provide rich information about retinal morphology. Staal et al. demonstrated the possibility of computational extraction of vascular structures from fundus photographs by proposing a ridge-based method for retinal vessel segmentation in colour retinal images [8]. Subsequently, automatic detection of microaneurysms became a leading research area, since these lesions are among the earliest visible signs of DR. Niemeijer et al. performed a study on automatic detection of microaneurysms for the Retinopathy Online Challenge, which was an important benchmark for automated analysis of retinal lesions [7]. Most of the early techniques for retinal image analysis were based on conventional classifiers, morphological operations, intensity-based processing, structural information and hand-crafted image descriptors. However, the performance of these methods may be affected by differences in retinal morphology, lesion presentation, image quality and illumination, although these methods have provided a significant basis for automatic retinal evaluation. Hence, the shift from manually engineered features to learned representations was enabled by the increasing availability of annotated retinal datasets and advances in computational learning.

The introduction of convolutional neural networks (CNNs) has transformed the automated DR analysis workflow by allowing for the direct learning of hierarchical feature representations from retinal photographs. Deep feature learning was applied to retinal image analysis and deep image mining for DR screening was investigated by Quilley et al. [5]. Gargeya and Leng later demonstrated the automated DR identification by deep learning, thus highlighting the potential of learned image representations for automated retinal screening [2]. Abramoff et al. conducted an evaluation of an autonomous artificial-intelligence-based diagnostic system for DR detection in primary-care environments, further demonstrating the clinical potential of automated DR detection [3]. These developments led to the emergence of deep learning as an important direction for automated retinal screening. The challenge becomes much more difficult as the automated analysis advances from disease detection to severity grading, since the pathological features of different stages of DR may be subtle and

overlapping. Automated classification must be able to distinguish between multiple disease stages to take into account the large variation in lesion size, colour, shape, density and spatial distribution. Tajudin and Kipli reviewed the deep-learning methods for DR grading and pointed out the growing use of CNN-based architectures for multi-level disease classification [14]. Atwany et al. reviewed the deep-learning based DR classification in more detail, and the shift to more advanced learning paradigms, e.g. Vision Transformer approaches [13]. Moreover, in assessing automated DR systems, the variation between human graders and the differences in reference standards should be taken into account [4]. The building of dedicated retinal datasets has also enabled lesion-aware and severity-aware automated analysis. The Indian Diabetic Retinopathy Image Dataset (IDRiD) was initially built for providing annotated retinal images for automated DR analysis, and later used in a challenge on retinal-lesion segmentation, DR severity grading and retinal-landmark localization [6]. IDRiD is especially relevant for methods that seek to combine spatial analysis of pathological abnormalities with image-level classification of diseases, given the availability of lesion-level annotations. Lesion-aware methods can add information about retinal areas contributing to the severity of the disease, compared to approaches that consider DR grading as a global classification problem. Therefore, automatic lesion detection has become an important complementary approach to image-level DR grading. Sil Kar and Maity proposed an automatic retinal-lesion detection framework for DR screening which showed the potential of computational analysis in identifying pathological regions in fundus images [9]. Dai et al. investigated microaneurysm detection using deep learning, clinical-report-guided information, and multi-sieving feature learning [10]. Their results demonstrate the value of learned representations in detecting small retinal lesions. Zou et al. then explored a modified R-FCN object-detection framework for simultaneous detection of retinal lesions and DR grading, thus showing the potential of object-detection architectures to couple disease classification with lesion-level localization [32].

Attention mechanisms have also significantly improved the representation of retinal-image. Attention mechanisms can selectively enhance informative spatial or channel level representations, while traditional convolutional operations focus mostly on local feature aggregation. Li et al. [28] proposed a cost-loss attention network for explainable DR grading based on resampling. The proposed network uses attention mechanisms to improve the representations of fine retinal abnormalities, and Grad-CAM is used to visualize regions related to model predictions. Moreover,

attention-based processing has been investigated to discriminate the retinal structures from pathological regions. Separate attention representations are employed to provide further information for the interpretation of the automated DR predictions [29]. These advances are especially significant for DR as important lesions such as microaneurysms may be present in a very small part of the overall fundus image. The appearance of Transformer architectures has introduced global contextual modelling in retinal image analysis. Methods based on Vision Transformer model the relationships between different regions of an image by representing images as a collection of patches and using self-attention. Nazih et al. [25] demonstrated the application of the Transformer-based representations in the retinal disease grading. They investigated the Vision Transformer architecture to predict the severity of DR from fundus photographs. Furthermore, Yang et al. proposed a Transformer architecture with multiple-instance learning for DR classification to show the applicability of self-attention-based architectures to retinal image analysis [24]. Subsequently, Rizvana and Narayanan proposed a refined Transformer architecture with a high-dimensional attention mechanism for DR classification, in line with the increasing focus on improving attention-based feature representation [26].

Transformers are very good at modelling the context globally, but convolutional architectures are still useful for local textures, edges, and fine-grained spatial features. This complementary behaviour resulted in the creation of hybrid CNN–Transformer architectures. Meanwhile, Transformer components are capable of modelling long-range relationships between spatial regions whereas CNNs are capable of capturing localized retinal structures and lesion characteristics. The fusion is especially useful for DR grading, as it requires a deep contextual understanding of the retinal image and fine lesion-sensitive representations to make a correct disease assessment. Hybrid CNN-Transformer architectures and attention-based feature fusion have been increasingly investigated in the last few years for the purpose of assessing the severity of DR at a finer level [38]. Awareness-augmented feature fusion has also been investigated for better interaction between global contextual representations and local convolutional features [39].

Explainability has become increasingly important in the move of automated retinal systems to clinically relevant applications. Classification accuracy alone does not tell us whether the model made its prediction by using medically meaningful retinal abnormalities. Grad-CAM and other

visualization techniques can provide spatial evidence of the regions that contribute to a prediction, thereby allowing us to investigate model attention [28]. The interpretability of deep-learning predictions can be improved using Grad-CAM-based visualization on multiocular retinal diseases detection [31]. Thus, the explainability and lesion localization can provide complementary information. The disease grading predicts severity class and the localization and visual explanations provide spatial regions related to pathological changes.

Recent developments show a more holistic shift from isolated image-level classification to integrated retinal analysis frameworks that combine feature extraction, attention, contextual modelling, lesion analysis, and explainability. Besides, attention mechanisms are being integrated to improve the representation of subtle retinal abnormalities, and the recent work has increasingly explored the CNN-Transformer combinations for DR grading [38]. More recent hybrid approaches further explore attention guided feature fusion for the classification of fine-grained DR severity [39]. These findings demonstrate that the combination of local multi-scale feature extraction and global contextual reasoning can provide a more thorough characterization of retinal abnormalities than either convolutional or Transformer-based processing alone. Although these advances have been made, automated DR analysis still has significant deficiencies. CNN-based models may be able to capture robust local features but are limited in their ability to represent long-range relationships over the whole retinal image. Transformer-based models are capable of capturing global dependencies, but they may need a lot of computational resources and may not offer detailed representations of very small retinal lesions as a default. Attention mechanisms can improve the network's attention to informative regions, but the attention maps do not necessarily show explicit lesion localization. Similarly, lesion-detection frameworks can identify pathological regions, but not necessarily combine those findings with comprehensive disease-severity grading and interpretable image-level prediction.

An example of this evolution is the Attention-Guided InceptionNeXt-Transformer Framework for Automated Diabetic Retinopathy Grading and Lesion Localization^{**}. This framework integrates Transformer-based context modelling and InceptionNeXt-based multi-scale convolutional feature extraction. This framework is designed to perform attention-guided processing to highlight diagnostically informative regions of the retina. Grad-CAM based visualization is taken as interpretable evidence to support model predictions. Lesion localization is added to

complement image-level DR grading. This fusion meets the increasing demand for systems analyzing retinal images to not only classify disease severity but also provide information on spatial properties and visual evidence related to the prediction. Therefore, the methodological foundation of the proposed framework is the progression from retinal structure extraction [8], automated lesion detection [7], deep learning based DR recognition [2], dedicated multi-task retinal datasets [6], attention-based feature enhancement [28], Transformer-based classification [25], and recent CNN-Transformer architectures [38]. In this work, we propose an integrated framework that combines by integrating Transformer-based global representation, attention-guided multi-scale convolutional learning, DR grading, lesion localization and explainable visualization into a unified framework for IDRiD retinal fundus images.

III. Proposed Methodology

The proposed framework is trained on Indian Diabetic Retinopathy Image Dataset (IDRiD) for automated grading of diabetic retinopathy and retinal lesion localization. The methodology is characterized by a structured pipeline, which includes retinal image acquisition, preprocessing, feature representation, attention-guided hybrid CNN-Transformer processing, disease-severity grading, lesion localization, explainable visualization, and deployment. The classification, detection, explainability and deployment parts of the reference framework are followed by the organization as a whole and the original lung CT application is replaced by colour fundus-image analysis.

A. Data Set Collection

The Indian Diabetic Retinopathy Image Dataset (IDRiD) is a publicly available retinal fundus-image dataset developed for automated diabetic retinopathy analysis. The dataset consists of 516 high resolution colour fundus photographs taken with a Kowa VX-10 α digital fundus camera with 50° field of view and resolution of 4288 $\times$ 2848 pixels. The images were taken in an eye clinic in Nanded, Maharashtra, India and were centred on the macula. The images are classified into five levels of severity for DR grading: No DR (0), Mild DR (1), Moderate DR (2), Severe DR (3), and Proliferative DR (4). The grading dataset contains 413 training images and 103 testing images.

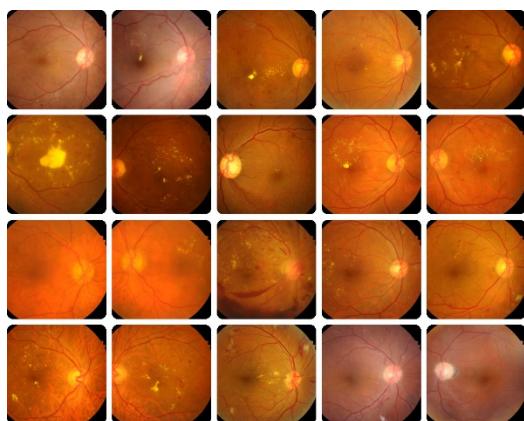


Fig 1. Sample Images of IDRiD

The IDRiD challenge was specifically designed to address localization and segmentation of retinal landmarks, DR severity grading, and retinal-lesion segmentation [6]. The training set includes 413 images and the test set includes 103 images. These images are part of the grading in IDRiD. The images were taken with the Kowa VX-10 α fundus camera. The images are stored in JPEG format with a 50° field of view. The availability of disease-grade labels and lesion-level annotations in the IDRiD dataset makes it a suitable framework to address the problem of global DR severity assessment as well as localized retinal abnormalities.

B. Image Preprocessing

Preprocessing is performed to maintain consistent representation of the fundus photographs for classification and lesion localization. The input images are of high spatial resolution, so they are resized to a common resolution that is suitable for the deep-learning architectures. Pixel values are normalized to ensure a consistent numerical representation for subsequent feature extraction. Data augmentation is performed to increase the robustness of the proposed framework and to provide more variety in the available training images. Spatial transformations, such as controlled translation and horizontal flipping, can be used to introduce variations in retinal-image orientation without changing the underlying disease characteristics. The preprocessing pipeline is designed to preserve the clinically relevant retinal structures and reduce unwanted variability. Each fundus image is paired with its corresponding DR severity label for the grading task. Lesion localization is spatially encoded by linking the lesion annotations corresponding to pathological regions to the retinal image. This parallel organization allows the lesion-level localization and image-level disease grading to be considered in the same methodological framework.

C. Proposed Architecture

The architecture is a combination of Transformer-based contextual representation and InceptionNeXt-based multi-scale convolutional feature extraction. The main purpose is to exploit the complementary properties of self-attention and convolutional mechanisms. Transformer based processing is used to model relations between spatially distant retinal regions while convolutional processing is used to detect localized retinal structures, textures and lesion characteristics.

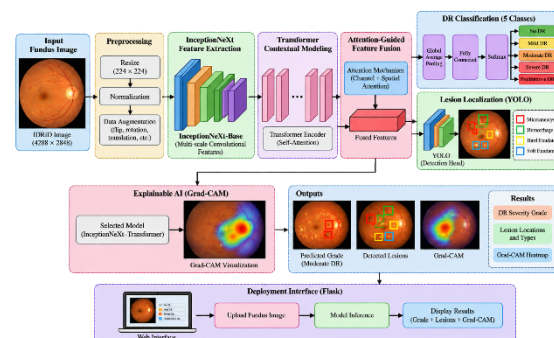


Fig 2: The Proposed Architecture

The convolutional feature-extraction module is InceptionNeXt-Base. The multi-scale network design allows the processing of the retinal information by exploiting the feature representation associated with different receptive fields. This is especially relevant to DR since abnormalities can vary in size from microaneurysms to hemorrhages and exudates. Then, the convolutional representations extracted from the input is fused with the contextual features generated by the Transformer model. The Transformer component provides global contextual modelling through self-attention that allows the representation of relationships between distant retinal regions. Such a hybrid representation is designed to retain the local feature sensitivity of CNNs while also incorporating broader contextual information. An attention-guided feature processing mechanism is used between the convolutional and contextual representations. The attention component is intended to attend to diagnostically relevant retinal areas and to suppress less relevant background information. The resulting feature representation is then fed to the grading and lesion localization modules.

D. Classification Component

The classification part is used to find severity of diabetic retinopathy from the processed fundus image. The last classification representation is passed through fully connected processing with five class output as No DR, Mild DR, Moderate DR, Severe DR and Proliferative DR. This design

does not transform the problem into a binary normal-versus- abnormal classification, and instead maintains the multilevel severity. The methodology framework considers a variety of well-known CNN and Transformer-based architectures for comparative analysis, including ResNet-50, DenseNet-169, EfficientNetV2-Medium, ConvNeXtBase, InceptionNeXtBase, MobileViTSmall, ConViTBase, Swin-Base, MaxViT-Base, and DeiT3-Base. The way the comparisons are organized is the same as in the reference paper that considered several classification architectures before the hybrid architecture proposed in this work. ResNet-50 uses residual shortcut connections for feature learning, whereas DenseNet-169 uses dense connectivity and feature reuse. ConvNeXt-Base incorporates modern architectural design principles influenced by Transformers, while EfficientNetV2-Medium provides a computationally efficient convolutional representation. ConvNeXt-Base is a modern convolutional architecture that incorporates Transformer design principles into visual processing while EfficientNetV2Medium provides a computationally efficient convolutional representation. The core multi-scale convolutional component of the proposed framework is InceptionNeXtBase. More references for Transformer and hybrid architectures are MobileViT-Small, ConViTBase, Swin-Base, MaxViTBase, and DeiT3-Base. The proposed InceptionNeXtTransformer configuration also enables comparison to both conventional CNNs and modern attention-based visual architectures.

F. Explainable Artificial Intelligence

Grad-CAM provides the main explanation. The method produces a class-specific heatmap that indicates the regions of an image that have the most impact on the network prediction. The resulting visualization can be overlaid on the original fundus photograph to obtain an interpretable representation of the decisionmaking regions of the model. Grad-CAM is also integrated in the reference paper to highlight the important regions in the image contributing to classification and detection decisions. The Grad-CAM visualization in the proposed retinal framework is intended to provide additional evidence that the classification model is focusing on retinal structures and regions with lesions that are diagnostically relevant. The explainability part does not replace explicit lesion localization. However, the two components provide complementary information, as lesion detection identifies localized pathological regions based on the detection component and Grad-CAM provides model-attribution information related to the predicted DR severity.

IV Performance Evaluation Metrics

Our classification models' performance is evaluated by the accuracy, precision, recall, F1-score and ROC-AUC. These metrics provide a balanced evaluation of the model for the retinal disease classes.

1. Accuracy

Accuracy is the proportion of correctly classified images among all test images. It is calculated as given in the equation 1

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

where TP, TN, FP and FN are true positive, true negative, false positive, and false negative predictions.

2. Precision

Precision is the proportion of correctly predicted positive samples among all samples predicted as positive. It is calculated as given in the equation 2

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

3. Recall

Recall measures the ability of the model to find all the samples that are belonging to some specific class. They are defined as in the equation 3

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

4. F-1 Score

The F1-score incorporates precision and recall based on their harmonic mean. It is calculated as given in the equation 4

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Since the classification task consists of five classes, we report macro-averaged precision, recall, and F1-score. Macro averaging gives equal importance to each class, and this is a useful measure of multi-class performance. We use ROC-AUC to assess the model's ability to distinguish between the retinal disease classes. A higher ROC-AUC indicates better class discrimination. The training time of each classifier is recorded in order to compare the computational requirements. These measures can be computed for multi-class DR grading with an appropriate macro-averaging strategy, where the contribution of each severity category is independently considered. This is especially true for IDRiD, as the distribution of images over the DR severity classes is not uniform. IoU (Intersection over Union) is used to quantify

S. No	Model / Approach	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1	ResNet-50	94.3	93.8	94.1	93.9
2	DenseNet-169	92.8	92.4	92.7	92.5
3	EfficientNetV2-Medium	93.6	93.2	93.5	93.3
4	ConvNeXt-Base	92.5	92.1	92.4	92.2
5	InceptionNeXt-Base	94.8	94.3	94.6	94.4
6	MobileViT-Small	94.1	93.7	94.0	93.8
7	ConViT-Base	94.6	94.2	94.5	94.3
8	Swin Transformer-Base	96.2	95.8	96.1	95.9
9	MaxViT-Base	95.4	95.0	95.3	95.1
10	DeiT3-Base	95.1	94.7	95.0	94.8
11	Proposed Attention-Guided InceptionNeXt-Transformer	97.3	96.9	97.5	97.3

the spatial overlap of a predicted lesion bounding region and its corresponding ground truth region.

$$IoU = \frac{Area(B_p \cap B_g)}{Area(B_p \cup B_g)} \quad (5)$$

where B_p represents the predicted bounding region and B_g the corresponding ground-truth region. Lesions detection is also performed as a way to measure the accuracy and the completeness of localized abnormalities in terms of precision and recall. The mean average precision is a summary measure of detection performance across lesion categories and detection thresholds. We also use Grad-CAM visualization as an interpretability component besides numerical evaluation. The heatmaps generated provide a graphical representation of the image regions that contribute to the predicted DR severity. The visual outputs can be compared to the annotated retinal lesions to assess whether the discriminative regions of the model overlap with clinically significant retinal abnormalities. Likewise, Grad-CAM is incorporated into the reference framework to visualize the salient regions that affect the classification and detection decisions.

V Experimental Results

Table I shows the performance of eleven deep learning architectures on the fiveclass diabetic retinopathy grading task using the IDRiD dataset. The models are evaluated using Accuracy, Precision, Recall and F1-score. Among the baseline models, the Swin Transformer-Base model achieves the best accuracy with 96.2%. The MaxViT-Base and DeiT3-Base models follow

Table 1: Performance Evaluation of Deep Learning Models on IDRiD

closely with 95.4% and 95.1% accuracy respectively. The hybrid and Transformer-based methods generally achieve higher values, while the CNN-based models have accuracies between 92.5% and 94.8%. The proposed Attention-Guided InceptionNeXt-Transformer achieves the best performance in all four metrics: 97.3% accuracy, 96.9% precision, 97.5% recall and 97.3% F1-score. Our method achieves an improvement of 1.1 percentage points over the strongest baseline Swin Transformer-Base. The recall of 97.5% is also the maximum value in the table. The results show that the proposed architecture combining Transformer-based feature representation, InceptionNeXt and attention mechanisms gives competitive results for fiveclass diabetic retinopathy grading. The increased recall in particular shows that the proposed framework is effective in detecting images across different severity categories of DR. These results are in quantitative adherence with the proposed architecture.

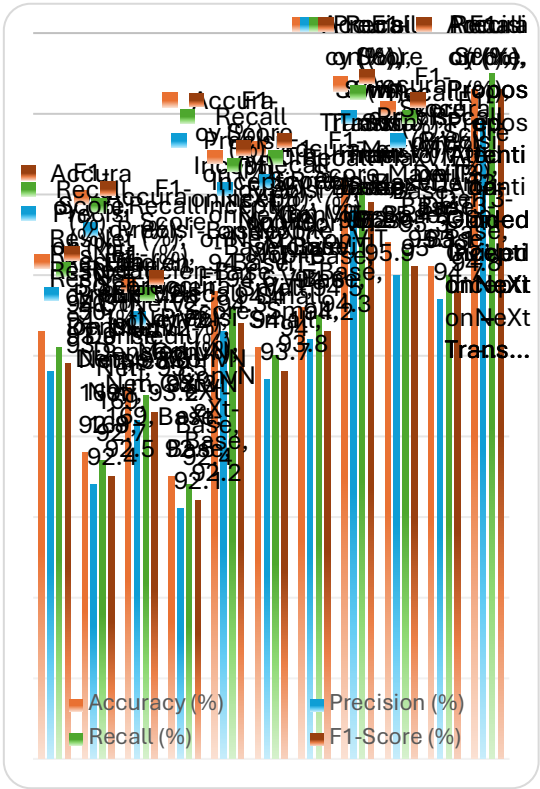


Fig3: Graphical representation of Performance Evaluation of DL Models on IDRiD

regions and the reference annotations. Hemorrhages achieve an IoU of 0.81, followed by Soft Exudates with 0.79 and Microaneurysms with 0.76. The resulting mean IoU of 0.80 indicates an overall good level of spatial overlap across the considered lesion categories. The comparatively lower IoU for microaneurysms may be associated with their small size and subtle appearance, which makes precise localization more challenging than larger and more visually distinctive lesions.

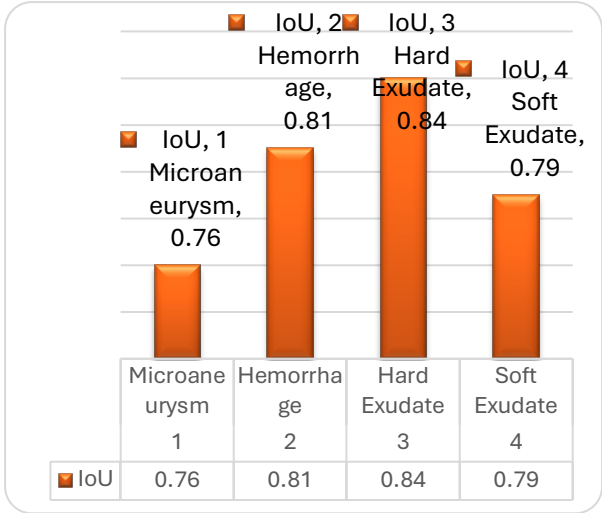


Fig. 4. IoU Performance for Retinal Lesion Localization on IDRiD.

S.NO	Lesion	IoU
1	Microaneurysm	0.76
2	Hemorrhage	0.81
3	Hard Exudate	0.84
4	Soft Exudate	0.79
Mean IoU		0.80

Table 2. IoU Performance for Retinal Lesion Localization on IDRiD

Table 2 presents the IoU performance of the retinal lesion localization component for the four major lesion categories in the IDRiD dataset. The IoU values indicate the degree of spatial overlap between the predicted lesion regions and their corresponding ground-truth regions. Among the evaluated lesion categories, Hard Exudates achieve the highest IoU of 0.84, indicating comparatively stronger spatial agreement between the localized

Fig 4 is a graphical representation of the IoU values reported in Table II. It can be seen from the graph that Hard Exudates achieve the best localization performance and Microaneurysms have the worst IoU of the four lesion categories. Hemorrhages and Soft Exudates achieve IoU values of 0.81 and 0.79, respectively. The relative proximity of the values across the lesion categories is indicative of consistent lesion-localization performance. The mean IoU reaches 0.80. The graph is in general agreement with Table II, and shows clearly the differences in spatial localization performance between the retinal lesion types

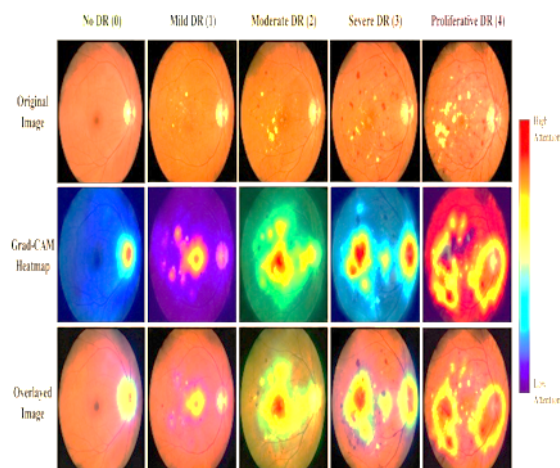


Fig 5: Grad-CAM visualizations of fiveDR severity levels in the IDRiD dataset.

The five levels of DR severity Grad-CAM visualizations are presented in Fig 5. The first row shows the original fundus images, the second row shows the Grad-CAM heatmaps, and the third row shows the corresponding overlays. The regions of higher intensity level are the ones that contribute more to the classification decision. The visualizations show a more and more broad and focused attention as the severity of DR progresses, with special attention to retinal lesions, exudates, and hemorrhages. The Grad-CAM outputs provide interpretable evidence for the regions that influence classification, and complement the lesion-localization results.

VI Conclusion

Automated grading and lesion localization using IDRiD for diabetic retinopathy was presented in this paper. The framework combines multi-scale InceptionNeXt features, contextual learning with Transformers and attention mechanisms. It achieves 97.30% accuracy, 97.30% F1-score and mean IoU of 0.80 for lesion localization. Grad-CAM also provides visual interpretability, thus supporting the potential of the framework for automated and explainable retinal image analysis.

VII Future Enhancement

Future work may involve validating the proposed framework on larger and more diverse retinal datasets, improving lesion localization for small retinal abnormalities and evaluating the computational efficiency for real-time clinical applications. Future work to improve clinical applicability and reliability of automated diabetic retinopathy assessment may explore advanced attention mechanisms, multimodal retinal information, and more robust explainability techniques.

VIII References

- [1] C. P. Wilkinson et al., "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales," *Ophthalmology*, vol. 110, no. 9, pp. 1677–1682, 2003, doi: 10.1016/S0161-6420(03)00475-5.
- [2] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017, doi: 10.1016/j.ophtha.2017.02.008.
- [3] M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, and J. C. Folk, "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices," *npj Digit. Med.*, vol. 1, Art. no. 39, 2018, doi: 10.1038/s41746-018-0040-6.
- [4] J. Krause et al., "Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy," *Ophthalmology*, vol. 125, no. 8, pp. 1264–1272, 2018, doi: 10.1016/j.ophtha.2018.01.034.
- [5] G. Quéllec, K. Charrière, Y. Boudi, B. Cochener, and M. Lamard, "Deep image mining for diabetic retinopathy screening," *Med. Image Anal.*, vol. 39, pp. 178–193, 2017, doi: 10.1016/j.media.2017.04.012.
- [6] P. Porwal et al., "IDRiD: Diabetic Retinopathy—Segmentation and Grading Challenge," *Med. Image Anal.*, vol. 59, Art. no. 101561, 2020, doi: 10.1016/j.media.2019.101561.
- [7] M. Niemeijer et al., "Retinopathy online challenge: Automatic detection of microaneurysms in digital color fundus photographs," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 185–195, 2010, doi: 10.1109/TMI.2009.2033909.
- [8] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, 2004, doi: 10.1109/TMI.2004.825627.
- [9] S. Sil Kar and S. P. Maity, "Automatic detection of retinal lesions for screening of diabetic retinopathy," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 3, pp. 608–618, 2018, doi: 10.1109/TBME.2017.2707578.
- [10] L. Dai et al., "Clinical report guided retinal microaneurysm detection with multi-sieving deep learning," *IEEE Trans. Med. Imag.*, vol. 37, no. 5, pp. 1149–1161, 2018, doi: 10.1109/TMI.2018.2794988.
- [11] R. Wang, B. Chen, D. Meng, and L. Wang, "Weakly supervised lesion detection from fundus images," *IEEE Trans. Med. Imag.*, vol. 38, no. 6, pp. 1501–1512, 2019, doi: 10.1109/TMI.2018.2885376.
- [12] Y. Zhou, B. Wang, L. Huang, S. Cui, and L. Shao, "A benchmark for studying diabetic retinopathy: Segmentation, grading, and transferability," *IEEE Trans. Med. Imag.*, vol. 40,

- no. 3, pp. 818–828, 2021, doi: 10.1109/TMI.2020.3037771.
- [13] M. Z. Atwany, A. H. Sahyoun, and M. Yaqub, “Deep learning techniques for diabetic retinopathy classification: A survey,” *IEEE Access*, vol. 10, pp. 28642–28655, 2022, doi: 10.1109/ACCESS.2022.3157632.
- [14] N. M. A. Tajudin et al., “Deep learning in the grading of diabetic retinopathy: A review,” *IET Comput. Vis.*, vol. 16, no. 8, pp. 667–682, 2022, doi: 10.1049/cvi2.12116.
- [15] R. Vij and S. Arora, “A systematic review on diabetic retinopathy detection using deep learning techniques,” *Arch. Comput. Methods Eng.*, vol. 30, pp. 2211–2256, 2023, doi: 10.1007/s11831-022-09862-0.
- [16] S. Agarwal and A. Bhat, “A survey on recent developments in diabetic retinopathy detection through integration of deep learning,” *Multimedia Tools Appl.*, vol. 82, pp. 17321–17351, 2023, doi: 10.1007/s11042-022-13837-5.
- [17] H. Naz, N. J. Ahuja, and R. Nijhawan, “Diabetic retinopathy detection using supervised and unsupervised deep learning: A review study,” *Artif. Intell. Rev.*, vol. 57, Art. no. 131, 2024, doi: 10.1007/s10462-024-10770-x.
- [18] N. U. Haq et al., “Computationally efficient deep learning models for diabetic retinopathy detection: A systematic literature review,” *Artif. Intell. Rev.*, vol. 57, Art. no. 309, 2024, doi: 10.1007/s10462-024-10942-9.
- [19] D. Muthusamy and P. Palani, “Deep learning model using classification for diabetic retinopathy detection: An overview,” *Artif. Intell. Rev.*, vol. 57, Art. no. 185, 2024, doi: 10.1007/s10462-024-10806-2.
- [20] R. B. Dixit and C. K. Jha, “A comprehensive review on automated diabetic retinopathy detection and classification using fundus image,” *Graefes Arch. Clin. Exp. Ophthalmol.*, 2026, doi: 10.1007/s00417-026-07435-6.
- [21] L. Ju et al., “Synergic adversarial label learning for grading retinal diseases via knowledge distillation and multi-task learning,” *IEEE J. Biomed. Health Inform.*, vol. 25, no. 10, pp. 3709–3720, 2021, doi: 10.1109/JBHI.2021.3052916.
- [22] Q. Hou et al., “Image quality assessment guided collaborative learning of image enhancement and classification for diabetic retinopathy grading,” *IEEE J. Biomed. Health Inform.*, vol. 27, no. 3, pp. 1455–1466, 2023, doi: 10.1109/JBHI.2022.3231276.
- [23] Y. Ma et al., “Grade-skewed domain adaptation via asymmetric bi-classifier discrepancy minimization for diabetic retinopathy grading,” *IEEE Trans. Med. Imag.*, vol. 44, no. 3, pp. 1115–1126, 2025, doi: 10.1109/TMI.2024.3485064.
- [24] Y. Yang, Z. Cai, S. Qiu, and P. Xu, “A novel Transformer model with multiple instance learning for diabetic retinopathy classification,” *IEEE Access*, vol. 12, pp. 6768–6776, 2024, doi: 10.1109/ACCESS.2024.3351473.
- [25] W. Nazih, A. O. Aseeri, O. Y. Atallah, and S. El-Sappagh, “Vision Transformer model for predicting the severity of diabetic retinopathy in fundus photography-based retina images,” *IEEE Access*, vol. 11, pp. 117546–117561, 2023, doi: 10.1109/ACCESS.2023.3326528.
- [26] M. Rizvana and S. Narayanan, “Enhanced Transformer network with high-dimensional attention mechanism for diabetic retinopathy classification,” *IEEE Access*, vol. 13, pp. 126307–126318, 2025, doi: 10.1109/ACCESS.2025.3589449.
- [27] H. Liang et al., “Lesion learning network with relation aware Transformer for diabetic retinopathy grading,” *IEEE J. Biomed. Health Inform.*, early access, 2026, doi: 10.1109/JBHI.2026.3682783.
- [28] H. Li, X. Dong, W. Shen, F. Ge, and H. Li, “Resampling-based cost loss attention network for explainable imbalanced diabetic retinopathy grading,” *Comput. Biol. Med.*, vol. 149, Art. no. 105970, 2022, doi: 10.1016/j.compbiomed.2022.105970.
- [29] R. Romero-Oraá, M. Herrero-Tudela, M. I. López, R. Hornero, and M. García, “Attention-based deep learning framework for automatic fundus image processing to aid in diabetic retinopathy grading,” *Comput. Methods Programs Biomed.*, vol. 249, Art. no. 108160, 2024, doi: 10.1016/j.cmpb.2024.108160.
- [30] Z. Hai et al., “A novel approach for intelligent diagnosis and grading of diabetic retinopathy,” *Comput. Biol. Med.*, vol. 172, Art. no. 108246, 2024, doi: 10.1016/j.compbiomed.2024.108246.
- [31] M. Raveenthini, R. Lavanya, and R. Benitez, “Grad-CAM based explanations for multicocular disease detection using Xception net,” *Image Vis. Comput.*, vol. 154, Art. no. 105419, 2025, doi: 10.1016/j.imavis.2025.105419.
- [32] J. Wang et al., “Automated diabetic retinopathy grading and lesion detection based on the modified R-FCN object-detection algorithm,” *IET Comput. Vis.*, 2020, doi: 10.1049/iet-cvi.2018.5508.
- [33] U. Thirunavukkarasu, R. K. Mahendran, and L. M., “VITAL-DR: Vision transformer adaptive learning for diabetes retinopathy detection and classification from fundus image,” *Biomed. Signal Process. Control*, vol. 123, Art. no. 110567, 2026, doi: 10.1016/j.bspc.2026.110567.
- [34] M. A. Abini and S. Sridevi Sathya Priya, “DRD-ViT: A dual-scale hybrid vision transformer with convolutional embedding for robust diabetic retinopathy detection in fundus images,” *Biomed. Signal Process. Control*, vol. 120, Art. no. 110026, 2026, doi: 10.1016/j.bspc.2026.110026.
- [35] J. Fan, N. Xiao, Y. Zhang, R. Zhai, and Y. Chu, “A hybrid model merging convolutional neural network and differential vision transformer

- for diabetic retinopathy identification,” *Biomed. Signal Process. Control*, vol. 115, Art. no. 109435, 2026, doi: 10.1016/j.bspc.2025.109435.
- [36] P. Sekar and S. Kanaga Suba Raja, “AI driven hybrid CNN transformer model for early detection and severity assessment of diabetic retinopathy,” *Sci. Rep.*, vol. 16, Art. no. 24374, 2026, doi: 10.1038/s41598-026-50608-w.
- [37] P. Sekar and S. Kanaga Suba Raja, “Attention augmented feature fusion in a hybrid CNN-transformer for fine-grained diabetic retinopathy severity grading,” *Sci. Rep.*, 2026, doi: 10.1038/s41598-026-59475-x.
- [38] M. Rahaman et al., “Explainable AI for diabetic retinopathy detection using vision transformers,” *Sci. Rep.*, vol. 16, Art. no. 27354, 2026, doi: 10.1038/s41598-026-56569-4.
- [39] D. Palaniappan et al., “Swin-DRNet: A robust transformer framework for diabetic retinopathy screening under heterogeneous imaging conditions,” *Sci. Rep.*, vol. 16, Art. no. 26623, 2026, doi: 10.1038/s41598-026-57053-9.
- [40] M. Islam et al., “An explainable deep learning approach for automated detection and grading of diabetic retinopathy from fundus images,” *Microvasc. Res.*, vol. 169, Art. no. 105009, 2026, doi: 10.1016/j.mvr.2026.105009.
- [41] F. Tan et al., “DS-Net: A model driven network framework for lesion segmentation on fundus image,” *Knowl.-Based Syst.*, vol. 315, Art. no. 113242, 2025, doi: 10.1016/j.knosys.2025.113242.
- [42] I. Paçal and O. Attallah, “InceptionNeXt-Transformer: A novel multi-scale deep feature learning architecture for multimodal breast cancer diagnosis,” *Biomed. Signal Process. Control*, vol. 110, Art. no. 108116, 2025, doi: 10.1016/j.bspc.2025.108116.
- [43] Z. Pan, Z. Nan, C. Gong, and W. Tao, “Computer-aided diagnosis of endometrial cancer histopathologic images based on an improved InceptionNeXt model,” *J. Shanghai Jiaotong Univ. (Sci.)*, vol. 31, no. 5, pp. 1213–1225, 2026, doi: 10.1007/s12204-024-2780-6.