www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 18(1), 2027, 332-353. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2027.v18i1.1532

Progressive Transfer Learning in Hybrid CNN-Transformer Architectures for Robust Diabetic Retinopathy Grading

Diego Alarcón-Paredes[1], Gustavo Alonso-Silverio[2],*, Iris Paola Guzmán-Guzmán[3], Elías Ventura-Molina[4],* and Antonio Alarcón-Paredes[5],*

[1]Centro de Innovación, Competitividad y Sostenibilidad, Universidad Autónoma de Guerrero, 39640, Mexico; dalarcon@uagro.mx

[2]Facultad de Ingeniería, Universidad Autónoma de Guerrero, 39087, Mexico; gsilverio@uagro.mx

[3]Facultad de Ciencias Químico-Biológicas, Universidad Autónoma de Guerrero, 39087, Mexico; ipguzman2@gmail.com

[4]Centro de Innovación y Desarrollo Tecnológico en Cómputo, Instituto Politécnico Nacional, Mexico City, 07738, Mexico; eventuram@ipn.mx

[5]Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico City, 07738, Mexico; aalarcon@cic.ipn.mx

* Corresponding authors: GA: gsilverio@uagro.mx, EV: eventuram@ipn.mx, AA: aalarcon@cic.ipn.mx

Abstract. Automated diabetic retinopathy (DR) grading requires both local feature extraction and global spatial reasoning. While Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) struggle individually with spatial context and data limitations respectively, we propose an optimized hybrid framework (ProGrad-ViT) combining an InceptionV3 backbone with a ViT encoder. To mitigate ViT overfitting on imbalanced data, we introduce a Progressive Transfer Learning regime, pre-training on a binary task before transferring to 5-class grading, and utilizing algorithmic class-weighting instead of synthetic oversampling. Trained on a composite dataset (APTOS/MESSIDOR) and independently tested on IDRiD, our model achieved a Quadratic Weighted Kappa (QWK) of 0.9048 and a Balanced Accuracy of 0.7534 (95% CI 0.7073–0.7996). Notably, Mild DR sensitivity reached 80.0% (95% CI 60.9–91.1%) with 99.4% healthy specificity. These results indicate stable behaviour under a cross-dataset protocol and support ProGrad-ViT as a candidate for automated multi-class DR triage. Looking towards clinical integration, prospective validation on consecutive and unselected patients remains necessary.

Keywords: Diabetic Retinopathy; Deep Learning; Vision Transformers; Progressive Transfer Learning; Hybrid Neural Networks; Medical Image Classification; Artificial Intelligence in Medicine; Automated Triage.

Article information
Received: Aug 6, 2026
Accepted: Sep 6, 2026

1 Introduction

Diabetes mellitus is a major global health concern, with an estimated 589 million individuals aged 20–79 currently affected, and projections indicating that the prevalence will rise to 853 million cases by 2050 (International Diabetes Federation, 2025). In 2024, diabetes was attributed to 3.4 million deaths globally, corresponding to one death every nine seconds (International Diabetes Federation, 2025). The methodology underlying these global estimates is described by Guariguata et al. (2011), and their impact on the ageing population is analysed by Sinclair et al. (2020). As a leading cause of vision loss and blindness, diabetic retinopathy (DR) represents one of the most serious microvascular complications of diabetes, particularly among the working-age population worldwide (Cho et al., 2018; International Diabetes Federation, 2013). Epidemiological studies highlight the clinical urgency of this condition, with approximately one-third of all individuals with diabetes presenting signs of DR (Khanna et al., 2023).

The pathological progression of DR is characterised by microvascular damage to the retina, a metabolic disorder caused by elevated blood sugar levels. Sustained hyperglycaemia disrupts the normal function of blood vessels, leading to fluid leakage, haemorrhages, and ultimately, retinal detachment and irreversible vision loss if not treated promptly (Watkins, 2003). As DR advances, specific lesions such as microaneurysms (MAs), haemorrhages (HMs), hard exudates (HEs), and neovascularisation begin to form. These lesions are recognised as critical biomarkers in determining the severity of the disease (Abràmoff et al., 2010). The International Clinical DR Classification System (Wilkinson et al., 2003) categorises the clinical diagnosis of DR into five distinct levels: no-DR, mild, moderate, severe non-proliferative DR (NPDR), and proliferative DR (PDR). Figure 1 illustrates the five classes of DR according to this scale.

Figure 1. *Progression of Diabetic Retinopathy (DR) severity in colour fundus photography according to the International Clinical DR Classification System. The representative images illustrate the typical progression of clinical biomarkers: (a) No-DR (healthy retina); (b) Mild DR (presence of isolated microaneurysms); (c) Moderate DR (emergence of haemorrhages and hard exudates); (d) Severe DR (extensive haemorrhages and significant exudation); and (e) Proliferative DR (advanced stage characterised by neovascularisation and preretinal haemorrhage).*

Early detection and timely management are paramount to slowing down or preventing blindness, as nearly 90% of DR cases can be effectively treated if diagnosed at an early stage (Hagos & Kant, 2019). Standard clinical practice relies on a thorough eye examination using fundus photography, a rapid and non-invasive method that captures high-resolution images of the retina (Kwan & Fawzi, 2019; Lechner et al., 2017). Despite its effectiveness, manual grading is time-consuming and highly dependent on the expertise of trained ophthalmologists. The escalating prevalence of diabetes has underscored an urgent need for automated computer-aided systems. Historically, pure CNNs have dominated this field; however, they exhibit an inductive bias towards local feature extraction, limiting the global spatial reasoning needed to separate the ordinal stages of DR, such as distinguishing moderate from severe disease on the basis of lesion density. ViTs, on the other hand, offer an alternative through self-attention mechanisms, but they are notoriously data-hungry and prone to severe overfitting on the imbalanced datasets scenarios typically found in medicine. A further concern is methodological, since much of the current literature relies on synthetic physical oversampling techniques that distort retinal morphology, frequently lacking rigorous cross-dataset validation; both practices tend to overestimate clinical viability (Fong et al., 2004; He et al., 2021).

To address these methodological gaps, this study proposes a unified, hybrid deep learning framework (ProGrad-ViT) that optimally integrates a CNN backbone (InceptionV3) with a Vision Transformer (ViT) encoder. Moving away from highly parameterised ensembles and synthetic data manipulation, our methodology is built upon a Progressive Transfer Learning regime. The primary contributions of this work are three-fold:

- Progressive Domain Transfer: We demonstrate that pre-training the hybrid network on a binary screening task (Phase 1) to filter general ophthalmological noise, and subsequently transferring these learned weights to a 5-class ordinal grading task (Phase 2), significantly stabilises the network. This approach mitigates the data-hunger of the ViT and eradicates the catastrophic forgetting typically observed in baseline CNNs.
- Algorithmic Balancing over Morphological Distortion: We substitute physical image oversampling with algorithmic class-weighting. This preserves the pristine morphological integrity of the original fundus images, allowing the self-attention mechanisms to accurately identify subtle biomarkers, such as isolated microaneurysms, thereby boosting the sensitivity of the critically elusive Mild DR stage to 80.0%.
- Rigorous Cross-Dataset Generalisation: To ensure real-world clinical applicability, the model was trained on a massive composite dataset (APTOS 2019 and MESSIDOR) and blindly tested on an independent, strictly curated dataset (the original IDRiD). Our optimised hybrid architecture achieved a Quadratic Weighted Kappa (QWK) of 0.9048 and a Balanced Accuracy of 0.7534, a result that is competitive with current literature benchmarks. By maintaining a sensitivity above 60% across all advanced stages, the proposed model provides a stable and reproducible basis for automated triage, subject to the prospective clinical validation detailed in the discussion section.

The remainder of this paper is organised as follows. Section 2 reviews the most relevant work on DR grading, from classical machine learning pipelines to recent convolutional and Transformer-based architectures. In Section 3, the datasets, the preprocessing pipeline, the proposed ProGrad-ViT architecture and the Progressive Transfer Learning methodology are described. Experimental results, state-of-the-art comparisons, clinical implications, and study limitations are presented in Section 4. Finally, Section 5 concludes the paper and outlines directions for future research.

2 Related Work

2.1. DR Grading and Classical Machine Learning

DR grading aims to achieve an accurate and automated assessment of disease severity for enhancing diagnosis precision and efficiency. Several machine learning models have been developed for the classification of DR. These automated systems employ advanced methods to extract features and representative information from retinal fundus images (Bhardwaj et al., 2021). Multilayer Perceptron (MLP), Support Vector Machines (SVM), and Random Forest (RF) approaches have been widely utilised in the literature.

The study presented by Mazlan et al. (2020) proposed a methodology for the automated detection of MAs in retinal fundus images using a MLP algorithm. This method was evaluated on the E-Ophta database, which comprises 100 fundus images. The MLP classifier achieved an accuracy of 92.28%. Majumder and Kehtarnavaz (2021) introduced a multitask learning framework for DR grading that integrates classification and regression by utilising task-specific loss functions. The extracted features were fused and input into a MLP, which functioned as the core classifier for categorising the five grades of DR. The multitask framework was evaluated on the APTOS and EyePACS datasets, achieving weighted Kappa scores of 0.90 and 0.88, respectively. The study reported by Wang et al. (2016) introduced a novel method for retinal vessel segmentation in fundus images. Feature extraction was applied to each pixel in the images, and the features obtained were input into a SVM algorithm to further classify the image pixels into vessels and non-vessels. The proposed method was evaluated on the DRIVE retinal database, achieving an average accuracy of 95.64%. Li et al. (2019) developed an automated DR detection system in which SVM plays a central role as the primary classifier. They trained two deep convolutional neural networks (DCNNs) with fractional max-pooling. Subsequently, their prediction results were combined using SVM. To optimise the SVM's performance, hyperparameters were fine-tuned using teaching-learning-based optimisation (TLBO), attaining an accuracy of 86.17% in the EyePACS dataset. The study reported by Seoud et al. (2015) proposed an automatic grading system for DR, employing a RF algorithm to classify fundus images into four severity levels, and reported a classification accuracy of 74.1%. Later, the study presented by Chowdhury et al. (2019) proposed a computer-aided diagnosis (CAD) system that utilises a RF classifier to detect retinal abnormalities caused by DR and age-related macular degeneration (AMD), achieving an accuracy of 93.58%.

These methods have shown potential and effectiveness for DR grading. However, several limitations can hinder their performance in real-world applications. A primary challenge is the reliance on handcrafted features and prior knowledge in a specific domain. This dependence significantly reduces the adaptability of machine learning approaches, especially under domain shifts, caused by diverse image acquisition conditions, scanning protocols, or heterogeneous demographics. Another critical limitation is related to the shallow structure of machine learning models, which restricts the capability for hierarchical feature learning. Consequently, they are not suitable for extracting complex patterns in retinal lesions, such as microaneurysms, HMs, and exudates, particularly at the early stages of DR.

2.2. Deep Learning and Convolutional Neural Networks

Recently, deep learning algorithms have marked a significant breakthrough in medical image analysis, contributing to the improvement of DR grading. CNNs are complex deep learning models that have demonstrated advanced performance in a wide range of computer vision tasks, including feature extraction, semantic segmentation, and target detection.

Reguant et al. (2021) employed a CNN-based visualisation method to analyse extracted features from retinal lesions in fundus images, specifically MAs, HMs, and exudates. The study conducted a comparative evaluation of four CNN architectures: InceptionV3, ResNet50, InceptionResNet50, and Xception. Each model was independently trained on the EyePACS and DIARETDB1 datasets. The classification performance for DR grading yielded accuracy rates ranging from 89% to 95%, with Area Under the Receiver Operating Characteristic Curve (ROC-AUC) values between 95% and 98%, sensitivity ranging from 74% to 86%, and specificity between 93% and 97%. The study presented by Ashwini and Dash (2023) introduced a novel approach for DR grading, integrating multiresolution feature extraction based on the Discrete Wavelet Transform (DWT) with deep learning-based classification using CNNs. As a pre-processing step, Contrast Limited Adaptive Histogram

Equalisation (CLAHE) was applied to enhance the contrast level in fundus images. The proposed method was implemented using ResNet, DenseNet, Xception, and EfficientNet CNN architectures. Each model was trained and evaluated independently on the IDRiD, DDR, and EyePACS datasets, achieving classification accuracies of 90.07%, 96.20%, and 93.53%, respectively. Sambyal et al. (2020) presented a modified U-Net architecture based on the pre-trained CNN ResNet to perform semantic segmentation of MAs and hard exudates in retinal fundus images. The proposed model was trained on the Ophta database and further evaluated on the IDRiD dataset. For MAs segmentation, the model attained 99.98% accuracy, 99.88% sensitivity, 99.89% specificity, and a dice score of 0.9998. For exudate segmentation, the model achieved 99.98% accuracy, 99.88% sensitivity, 99.89% specificity, and a dice score of 0.9999. Later, the study presented by Skouta et al. (2022) introduced a customised CNN based on the U-Net architecture, designed explicitly for the semantic segmentation of retinal HMs in fundus images. The model was trained on the IDRiD dataset and evaluated on both the IDRiD and DIARETDB1 datasets. The proposed method effectively segmented the bleeding of the HMs, achieving 98.68% accuracy, 80.49% sensitivity, and 99.68% specificity. Furthermore, the model achieved an Intersection over Union (IoU) of 76.61% and a Dice value of 86.51%. Liu et al. (2021) introduced a novel method for the early detection of DR, based on a deep symmetric convolutional neural network, to identify MAs and HEs. The model was trained and evaluated on the DIARETDB1 dataset, exploring three filtering structures: convolution, max-pooling, and average-pooling, which achieved detection accuracies of 92.0%, 93.2%, and 93.6%, respectively. The study reported by Usman et al. (2023) proposed a deep learning-based approach for DR diagnosis, referred to as the multi-level feature extraction and classification (ML-FEC) framework. According to the methodology, a first pre-processing step was applied to the colour fundus photographs (CFPs). Secondly, feature extraction was performed utilising Principal Component Analysis (PCA). Furthermore, their method leverages pre-trained CNN architectures, including ResNet50, ResNet152, and SqueezeNet1, and then applies transfer learning. These models were trained on a new subset of the CFPs, achieving classification accuracies of 93.67%, 94.40%, and 91.94%, respectively.

The aforementioned studies have made significant contributions to the classification of DR. However, accurately grading DR remains a highly challenging task due to the complexity of retinal lesions: fundus images exhibit inter-class similarities and intra-class differences (Mansour, 2018). That is, lesions located in different classes may appear visually similar in shape and texture, while notable differences in lesion size may be present within the same class. As a result, the classification of DR severity levels heavily depends on the integration of diverse types of features. However, despite their success in detecting local anomalies, CNNs inherently suffer from a limited receptive field. This inductive bias restricts their ability to capture long-range dependencies and the global spatial context of the retina. This global perspective is critical for accurately distinguishing between closely related advanced DR stages, which often depend on evaluating the widespread distribution and density of lesions across the entire fundus rather than isolated textures.

2.3. Vision Transformers and Progressive Learning

To overcome the spatial limitations of CNNs, Vision Transformers (ViTs) have recently emerged as a powerful paradigm in medical image analysis (Azad et al., 2024). Unlike standard convolution operations, ViTs utilise self-attention mechanisms to process images as sequences of patches, enabling the model to capture global context and long-range dependencies from the very first layer (Vaswani et al., 2017). In the context of DR grading, this global receptive field allows the network to dynamically correlate scattered lesions, such as multiple haemorrhages or hard exudates located in different retinal quadrants, providing a more holistic and clinically accurate assessment of disease severity (Banerjee et al., 2026).

Despite their architectural advantages, ViTs inherently lack the strong inductive biases of CNNs, such as translation invariance, making them notoriously "data-hungry". Training ViTs from scratch on small or highly imbalanced medical datasets often leads to severe overfitting and suboptimal generalisation (Nazih et al., 2023). To mitigate this, recent studies have actively explored the application of pure ViTs and hybrid CNN-Transformer architectures for DR grading. For instance, Liu et al. (2025) proposed STMF-DRNet, a multi-branch cascading framework with a Swin-TransformerV2 backbone; when evaluated across multiple datasets including DDR, EyePACS, APTOS-2019, and a clinical cohort, the model demonstrated robust performance on clinical data, achieving an accuracy, recall, and F1-score of 0.77, alongside a specificity of 0.942 and a Kappa score of 0.877. Similarly, the study presented by Yi et al. (2026) introduced CTANet, a hybrid framework that dynamically integrates local convolutional features and global window-based ViT context via an Adaptive Focus Fusion module; this approach achieved classification accuracies of 90.08% and 85.64% on the APTOS 2019 and DDR datasets, respectively. Furthermore, the work by Ikram and Imran (2025) introduced ResViT FusionNet, integrating a ResNet50 backbone with a Vision Transformer, and employed physical data augmentation techniques to balance the APTOS 2019 dataset; this model achieved an accuracy of 0.9301, along with a precision of 0.9307, recall of 0.9300, F1-score of 0.9275, and a Kappa score of 0.8935. However, to compensate for data scarcity and extreme class imbalance, many of these state-of-the-art approaches still rely on synthetic data manipulation, such as physical oversampling. This practice can artificially distort

the natural morphological distribution of retinal biomarkers and compromise the model's reliability in real-world scenarios (Bhoyar & Patel, 2026).

Therefore, a significant clinical and technical challenge remains open: effectively harnessing the global spatial reasoning of ViTs while preventing overfitting in imbalanced datasets, without altering the pristine morphology of the original images. The present work addresses these key limitations by introducing a novel hybrid architecture (ProGrad-ViT) trained on a massive composite dataset. Moving away from rigid hierarchical ensembles, our approach leverages a Progressive Transfer Learning regime. By pre-training the model on a binary screening task and systematically transferring the learned weights to a multi-class ordinal grading task, our methodology stabilises the Transformer's attention mechanisms, mitigates class imbalance through algorithmic weighting, and dramatically enhances robustness and generalisability across diverse clinical settings.

3 Methodology

This section describes the datasets and the way they were integrated, the preprocessing pipeline, the proposed hybrid architecture, the two-phase Progressive Transfer Learning regime, the training configuration and the evaluation protocol, at the level of detail required to reproduce the experiments.

3.1. Dataset Integration and Cross-Validation Strategy

To train and evaluate the proposed progressive hybrid architecture, three publicly available benchmark datasets were utilised: APTOS 2019 (APTOS 2019 Blindness Detection, 2019), MESSIDOR (Decencière et al., 2014) and IDRiD (Porwal et al., 2018). A critical challenge in automated DR grading is the domain shift caused by varying image acquisition protocols, equipment, and patient demographics. To mitigate this limitation and enhance the model's robustness and generalisability, a composite training dataset was created by integrating the APTOS 2019 and MESSIDOR databases.

This combined dataset, comprising a total of 5,406 colour fundus images, was strictly partitioned into an 85% training set and a 15% validation set. This robust, heterogeneous training environment exposes the network to diverse retinal morphologies and lighting conditions. The detailed distribution of images across the five DR severity classes for both the training and validation subsets is presented in Table 1.

To ensure strict clinical viability and prevent data leakage, the Indian Diabetic Retinopathy Image Dataset (IDRiD) was maintained entirely isolated as an independent hold-out test set. The IDRiD dataset consists of 516 fundus images categorised into the five severity stages, providing a challenging and unseen environment for the model, as detailed in Table 2. Validating the architecture exclusively on this blind data guarantees an unbiased evaluation of its diagnostic performance and cross-dataset generalisation capabilities.

The partition was performed at the image level. Neither APTOS 2019 nor MESSIDOR distributes patient identifiers, so a patient-level split could not be enforced on the composite corpus. In MESSIDOR, which contains paired acquisitions for part of its population, the two eyes of the same individual may therefore fall on opposite sides of the training/validation boundary. This constraint does not compromise the results reported in Section 4, for two reasons. First, the validation subset is used exclusively for early stopping and learning-rate scheduling, and never to report diagnostic performance. Second, all reported metrics are obtained on IDRiD, a dataset acquired in a different country, with different equipment and by a different grading team, so no patient can appear simultaneously in the training material and in the evaluation material. Furthermore, a redundancy-screening procedure was performed, that is, exact duplicates were searched by MD5 checksum and near-duplicates with a 64-bit perceptual hash (Hamming distance ≤ 5), computed within each database, between the two databases, and crucially against the IDRiD hold-out set. This rigorous screening confirmed that there was zero data leakage, no exact or near-duplicate images were detected across the partitions, and therefore no images had to be removed.

Table 1. Description of the APTOS 2019 and the MESSIDOR datasets.

Dataset	Classes	Train images	Validation images	Raw images
APTOS 2019	01-No-DR	1535	270	1805
	02-Mild	315	55	370
	03-Moderate	850	149	999
	04-Severe	165	28	193
	05-Proliferative	250	45	295

	Subtotal	**3115**	**547**	**3662**
MESSIDOR	01-No-DR	866	151	1017
	02-Mild	230	40	270
	03-Moderate	296	51	347
	04-Severe	64	11	75
	05-Proliferative	30	5	35
	Subtotal	**1486**	**258**	**1744**
	Total	**4601**	**805**	**5406**

Table 2. Description of the IDRiD dataset.

Dataset	**Classes**	**Test images**	**Raw images**
IDRiD	01-No-DR	168	168
	02-Mild	25	25
	03-Moderate	168	168
	04-Severe	93	93
	05-Proliferative	62	62
	Total	**516**	**516**

3.2. Image Preprocessing and Algorithmic Class-Weighting

Retinal fundus images frequently suffer from uneven illumination, low contrast, and noise variations depending on the acquisition device. To standardise the input space and highlight critical clinical biomarkers, a rigorous preprocessing pipeline was applied to all images across the datasets. Initially, black background borders were cropped to centre the field of view on the retina, and all images were uniformly resized to a standard resolution of 512 × 512 pixels to match the input requirements of the hybrid network architecture.

Subsequently, Contrast Limited Adaptive Histogram Equalisation (CLAHE) was applied to the green channel of the images, as this channel exhibits the highest contrast for vascular structures and retinal lesions. The CLAHE algorithm enhances the local contrast of subtle biomarkers, such as isolated microaneurysms and hard exudates, while preventing the over-amplification of noise common in standard histogram equalisation. For this study, the CLAHE parameters were empirically set to a clip limit of 2.0 and a tile grid size of 8 × 8. Finally, all pixel values were normalised to a range of [0, 1] to accelerate network convergence.

A critical challenge in medical datasets is the inherent imbalance of disease severity classes, heavily skewed toward the No-DR and Mild stages. While recent state-of-the-art methodologies often rely on physical data augmentation or synthetic oversampling to balance class distributions, this study explicitly discarded these techniques. Physical oversampling risks artificially distorting the natural morphology and spatial distribution of the retina, which is detrimental to the global spatial reasoning of Vision Transformers. Instead, we implemented an algorithmic class-weighting strategy during the training phase. By calculating dynamic class weights derived from the inverse frequency of each stage in the training distribution, the loss function was heavily penalised for misclassifying minority classes (e.g., Severe and Proliferative DR). This approach ensures that the network prioritises learning robust features for underrepresented clinical stages without manipulating the pristine morphological integrity of the original fundus images.

3.3. Proposed Hybrid Architecture: ProGrad-ViT

To accurately grade the severity of Diabetic Retinopathy, a diagnostic system must simultaneously detect subtle, pixel-level anomalies (such as microaneurysms) and evaluate the widespread distribution of lesions across the entire retina. To fulfil this dual requirement, we propose the ProGrad-ViT architecture, a unified hybrid deep learning framework that seamlessly synergises the local feature extraction capabilities of Convolutional Neural Networks (CNNs) with the global spatial reasoning of Vision Transformers (ViTs).

3.3.1. CNN Backbone for Local Feature Extraction

The foundation of the ProGrad-ViT model relies on an InceptionV3 architecture acting as the primary feature extractor. InceptionV3 was selected due to its highly efficient multi-scale processing capabilities, driven by its internal inception modules. By utilising parallel convolutional filters of varying sizes within the same layer, the backbone can simultaneously capture pathological features of different dimensions. In the clinical context of DR, this structural advantage allows the network to effectively identify minuscule, isolated microaneurysms alongside larger, irregular hard exudates. The InceptionV3 backbone processes the 512 x 512 pixel fundus images and generates a rich, high-level spatial feature map, inherently maintaining the translation invariance that standard ViTs lack.

The selection of InceptionV3 over more recent backbones was deliberate and rests on three arguments. First, the factorised multi-branch design places parallel filters of different sizes inside every block, enabling multi-scale analysis at each network depth. Compound-scaled families such as EfficientNet, and modern large-kernel designs such as ConvNeXt, obtain a comparable diversity of receptive fields mainly by scaling depth and input resolution across the network rather than within the block. In DR, diagnostically decisive evidence spans two widely different scales simultaneously: microaneurysms a few pixels wide, and confluent haemorrhages or exudate plaques that occupy an entire quadrant. Per-block multi-scale filtering is therefore directly aligned with the task. Second, at the input resolution adopted here, the backbone yields a $14 \times 14 \times 2048$ feature map, equivalent to a 196-token sequence that matches the canonical ViT patch budget without additional pooling or interpolation. Hierarchical backbones such as Swin-V2, used in several recent hybrids, already incorporate windowed attention, which would duplicate the very component whose contribution this study sets out to isolate. Third, InceptionV3 requires approximately 23.9 M parameters, against roughly 88 M for ConvNeXt-B and a comparable or larger budget for the upper EfficientNet variants. This matters for the resource-limited telemedicine settings that motivate this work, and it also keeps the ablation study interpretable, since any gain observed can be attributed to the self-attention encoder and to the progressive regime rather than to a larger convolutional trunk. A systematic benchmark of alternative backbones under an identical progressive protocol remains a natural extension of this study.

3.3.2. Vision Transformer Encoder for Global Spatial Reasoning

While the CNN backbone excels at locating individual lesions, determining the precise ordinal stage of DR often depends on the density and spatial distribution of these lesions across multiple retinal quadrants. To overcome the restricted receptive field of convolutions, the high-level feature maps generated by the InceptionV3 backbone, yielding spatial dimensions of 14 x 14 with 2048 channels, are systematically transformed. First, the feature map is reshaped into a sequence of 196 discrete patches (14 x 14), and the channel depth is projected into a lower-dimensional embedding space of 256 via a dense layer to optimise computational efficiency.

This sequence is then fed into the Vision Transformer (ViT) encoder. The core component of the ViT, the Multi-Head Self-Attention (MHSA) mechanism, processes the sequence utilising 4 parallel attention heads (num_heads = 4) with a key dimension of 256. The encoder block also features a feed-forward network with a dimensionality of 256 (ff_dim = 256) and employs rigorous Layer Normalisation and residual connections to stabilise the gradients. Unlike standard convolutions, the MHSA dynamically computes the relevance of different retinal regions against one another, effectively correlating distant features regardless of their physical proximity in the image. To mitigate overfitting, a dropout rate of 0.3 is applied within the transformer block.

3.3.3. Classification Head

Following the self-attention blocks, the sequence of global representations is aggregated into a single, comprehensive feature vector using a 1D Global Average Pooling layer. To further ensure regularisation, a subsequent dropout layer with a rate of 0.5 is applied. Finally, a classification head utilising a softmax activation function outputs the predicted probability distribution across the five DR severity classes. This hybrid design ensures that the final clinical prediction is informed by both high-fidelity local textures and complex global context.

3.4. Progressive Transfer Learning Regime

Standard deep learning approaches often struggle to train hybrid CNN-Transformer architectures directly on multi-class, highly imbalanced medical datasets. Due to the lack of strong inductive biases in the self-attention layers, direct optimisation frequently leads to unstable attention maps and catastrophic forgetting of minority clinical stages. To overcome this limitation, we introduce a Progressive Transfer Learning regime, as illustrated in Figure 2. This strategy breaks down the complex 5-class ordinal grading task into a sequential, multi-phase learning curriculum, systematically guiding the network from general feature comprehension to fine-grained disease grading.

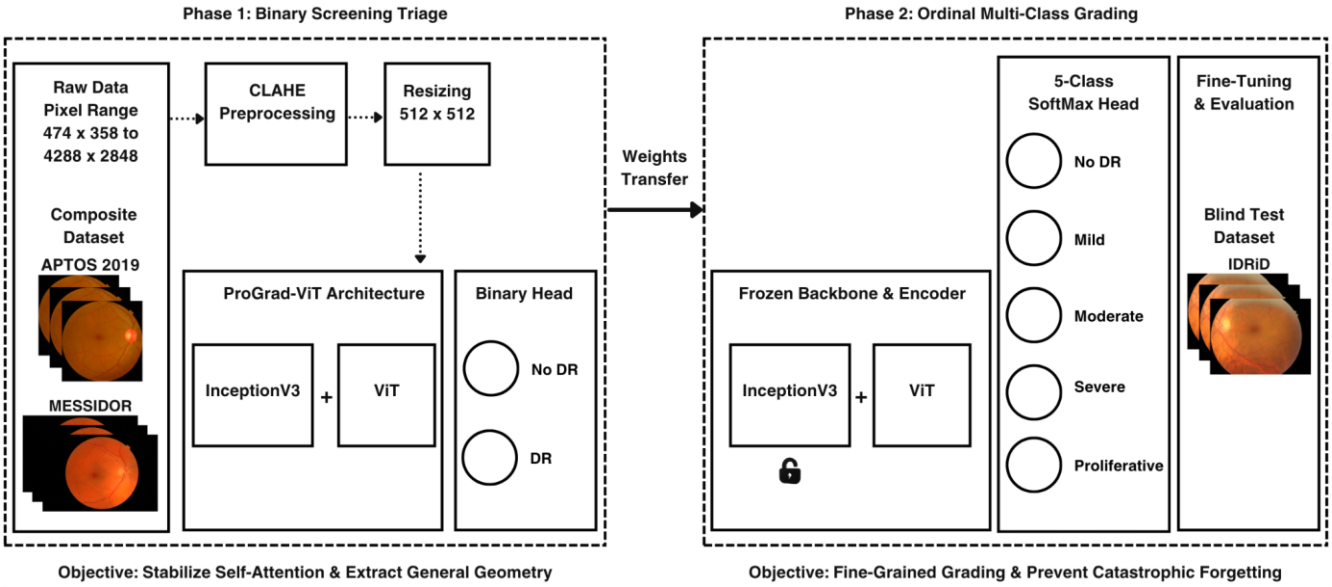

***Figure 2**. Conceptual flowchart of the proposed Progressive Transfer Learning regime. In Phase 1 (left), the ProGrad-ViT hybrid architecture is pre-trained on a binary screening task using a large composite dataset to stabilise the Vision Transformer's self-attention mechanisms and learn fundamental retinal morphologies. The optimised weights are subsequently transferred to Phase 2 (right), where the classification head is adapted for 5-class ordinal grading. This progressive adaptation isolates fine-grained pathological features and prevents the catastrophic forgetting of minority clinical stages.*

3.4.1. Phase 1: Binary Screening Triage (Knowledge Initialisation)

In the first phase, the hybrid network architecture is trained on a simplified binary classification task designed for screening triage. The objective is to differentiate strictly between healthy retinas (Class 01: No-DR) and diseased retinas (combining Classes 02 to 05 into a single "DR-Present" target).

During this phase, the network leverages the combined composite training set to learn fundamental ocular geometry, vascular structures, and general pathological anomalies without the burden of resolving complex ordinal boundaries. This binary optimisation acts as a critical regulariser, initialising and stabilising the MHSA weights of the Vision Transformer encoder on the medical domain. By resolving this macro-level task first, the network effectively satisfies the "data-hunger" of the ViT component, establishing a robust and domain-specific feature baseline.

3.4.2. Phase 2: Ordinal Multi-class Grading (Fine-Grained Adaptation)

Once Phase 1 achieves convergence, the learned weights from the entire hybrid architecture are extracted and transferred to initialise the final multi-class grading model. To adapt the network to the definitive diagnostic task, the binary classification head is replaced with a 5-class softmax layer corresponding to the complete International Clinical DR Classification System.

The training in Phase 2 employs a layered fine-tuning mechanism to ensure a smooth transition of domain knowledge without disrupting the structural features learned during the screening phase. Initially, the weights transferred from the InceptionV3 backbone and the ViT encoder are preserved, allowing the new 5-class classification head to optimise its dense connections. Subsequently, the entire hybrid network is unfrozen and optimised jointly using a highly conservative learning rate. This progressive transition enables the ProGrad-ViT model to focus its attention mechanisms on capturing the subtle, fine-grained visual differences that separate adjacent disease stages (such as the transition from Mild to Moderate NPDR) while completely safeguarding the network against catastrophic forgetting.

3.5. Training Details and Hyperparameters

The proposed architecture was implemented using the TensorFlow and Keras deep learning frameworks. To ensure reproducibility and optimal convergence, distinct hyperparameter configurations were carefully tailored for each phase of the Progressive Transfer Learning regime. Throughout all training stages, the categorical cross-entropy loss function was utilised, strictly coupled with the algorithmic class-weighting strategy to penalise the misclassification of minority instances.

3.5.1. Phase 1: Binary Screening Training Configuration

For the initialisation phase, the model was trained end-to-end to classify between healthy and diseased retinas. The network was optimised using the Adam optimiser with an initial learning rate of 1e-4. The training process was scheduled for a maximum of 30 epochs. To prevent overfitting, an Early Stopping callback was implemented with a patience of 8 epochs, monitoring the validation ROC-AUC to restore the best-performing weights. Additionally, a ReduceLROnPlateau callback was configured to halve the learning rate (factor of 0.5) if the validation loss plateaued for 5 consecutive epochs, down to a minimum learning rate of 1e-6.

3.5.2. Phase 2: Multi-class Fine-Tuning Configuration

The transition to the 5-class ordinal grading task required a meticulous, two-step fine-tuning approach to safeguard the weights inherited from Phase 1.

- Step 1 (Strategic Freezing): Initially, the InceptionV3 backbone was completely frozen (trainable = False), restricting gradient updates exclusively to the Vision Transformer encoder and the newly initialised 5-class dense layer. This step was optimised using Adam with a learning rate of 1e-4 and trained for 10 epochs. This strategic freezing allowed the classification head to adapt to the new ordinal boundaries without distorting the pre-trained feature extractor.
- Step 2 (Deep Fine-Tuning): Subsequently, the entire hybrid network was unfrozen (trainable = True) to allow joint optimisation of both local and global representations. To prevent catastrophic forgetting, the Adam optimiser's learning rate was drastically reduced to 1e-6. This deep fine-tuning phase was scheduled for 40 epochs. The regularisation constraints were also tightened: the Early Stopping patience was extended to 12 epochs (monitoring validation ROC-AUC), and the ReduceLROnPlateau callback was adjusted to a patience of 4 epochs with a minimum learning rate of 1e-8.

This meticulous scheduling ensures a smooth gradient descent across the highly complex loss landscape associated with multi-class, imbalanced medical datasets.

3.6. Evaluation Metrics

To rigorously evaluate the clinical viability and generalisation capabilities of the ProGrad-ViT architecture, a comprehensive suite of statistical metrics was derived from the confusion matrix. The fundamental evaluation is based on the calculation of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) for each respective clinical stage.

The standard metrics computed for the independent test set (IDRiD) include Accuracy, Sensitivity (Recall), and Specificity. These metrics are mathematically defined as follows:

$$Accuracy = \frac{TP \pm TN}{TP + TN + FP + FN} \tag{1}$$

$$Sensitivity\ (Recall) = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

Furthermore, the ROC-AUC was utilised to measure the model's capability to distinguish between classes across varying classification thresholds. This metric is calculated by integrating the True Positive Rate (TPR) over the False Positive Rate (FPR):

$$AUC = \int_0^1 TPR(FPR)d(FPR) \tag{4}$$

While standard overall Accuracy is widely reported in the state-of-the-art literature, and is included in this study to establish a baseline for comparative discussion, it can be a highly misleading metric in datasets with extreme natural class imbalances. To ensure that the model's performance was not artificially inflated by a bias toward majority classes (e.g., No-DR), the Balanced Accuracy was prioritised. This metric provides a transparent reflection of the network's diagnostic robustness across both underrepresented and majority clinical stages:

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} \tag{5}$$

Crucially, Diabetic Retinopathy grading is not a nominal classification problem, but an ordinal one; the severity of the pathology progresses linearly from Stage 01 to Stage 05. Consequently, the QWK score was selected as the definitive evaluation metric. Unlike standard multi-class accuracy, QWK heavily penalises misclassifications that are further away from the true ordinal label, aligning perfectly with clinical risk assessment protocols.

The QWK is calculated by first defining a weight matrix $w_{i,j}$ based on the squared difference between the actual and predicted ratings:

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2} \tag{6}$$

where i and j represent the true and predicted labels respectively, and N is the total number of classes. The final k score is then computed using the observed confusion matrix O and the expected (chance) matrix E:

$$k = 1 - \frac{\Sigma_{i,j} w_{i,j} O_{i,j}}{\Sigma_{i,j} w_{i,j} E_{i,j}} \tag{7}$$

3.6.1. Quantification of Statistical Uncertainty

The independent test set is small, and the 02-Mild class contains only 25 images, so every point estimate is reported together with a measure of its uncertainty. Class-wise sensitivity and specificity are accompanied by 95% Wilson score intervals, which remain well behaved for small samples and for proportions close to 0 or 1, where the normal approximation breaks down. The interval for balanced accuracy is derived by propagating the class-wise binomial variances: the five recalls are estimated on disjoint subsets of images and are therefore mutually independent, so the variance of their mean is the sum of the individual variances divided by the square of the number of classes. For the macro-average ROC-AUC and for the QWK, which are not simple proportions, 95% intervals are obtained by non-parametric bootstrap resampling of the test set, using 1,000 resamples stratified by true class and reporting the 2.5th and 97.5th percentiles of the resulting distributions. All intervals are given at the 95% level, and are truncated to the [0, 1] range wherever the analytical bound falls outside it.

By evaluating the ProGrad-ViT framework through this rigorous statistical validation, the reported results strictly reflect its true diagnostic safety and reliability.

4 Results and Discussion

4.1. Results

4.1.1. Phase 1: Binary Screening Performance (Knowledge Initialisation)

The primary objective of Phase 1 was to establish a robust feature baseline by training the hybrid architecture on a binary screening task to distinguish between healthy retinas (01-No-DR) and diseased retinas (DR). The evaluation on the strictly isolated IDRiD test dataset, comprising 516 images, demonstrated exceptional diagnostic precision.

The detailed class-wise clinical metrics are presented in Table 3. The Phase 1 model achieved a Balanced Accuracy of 0.9956, reflecting its symmetric capability to detect both healthy and pathological retinas without susceptibility to class imbalance.

Specifically, the network recorded a Sensitivity (Recall for DR instances) of 0.9971 and a Specificity (Recall for No-DR instances) of 0.9940.

Table 3. Clinical evaluation metrics for the Phase 1 Binary Screening task (InceptionV3 + ViT on the IDRiD hold-out test set).

Class	Accuracy	Sensitivity (Recall) [95% CI]	Specificity
01-No-DR	0.9940	0.9940 (0.9978 - 1.0000)	0.9971
DR	0.9971	0.9971 (0.9912 - 1.0000)	0.9940
Global Metrics Balanced Accuracy: 0.9956 (0.9874 - 1.0000) ROC-AUC: 1.0000 (0.9998 - 1.0000) QWK: 0.9912 (0.9775 - 1.0000)			

A granular visual analysis of the model's performance is illustrated in Figure 3. The learning curves (Figure 3a) show a stable and convergent training process, effectively minimising the categorical cross-entropy loss without signs of overfitting. The confusion matrix (Figure 3b) corroborates this reliability: out of the 516 blind test cases, the model produced only a single false positive and a single false negative, correctly identifying 167 out of 168 No-DR cases and 347 out of 348 DR cases.

Furthermore, the model exhibited an optimal class separability, as evidenced by the Receiver Operating Characteristic curve (Figure 3c) yielding a ROC-AUC score of 1.0000. To validate the stability of the learned features prior to transferring the weights, the QWK score was calculated, reaching an outstanding 0.9912. Clinically, this translates to successfully establishing a highly robust, non-overfitted initialisation for the subsequent multi-class grading phase.

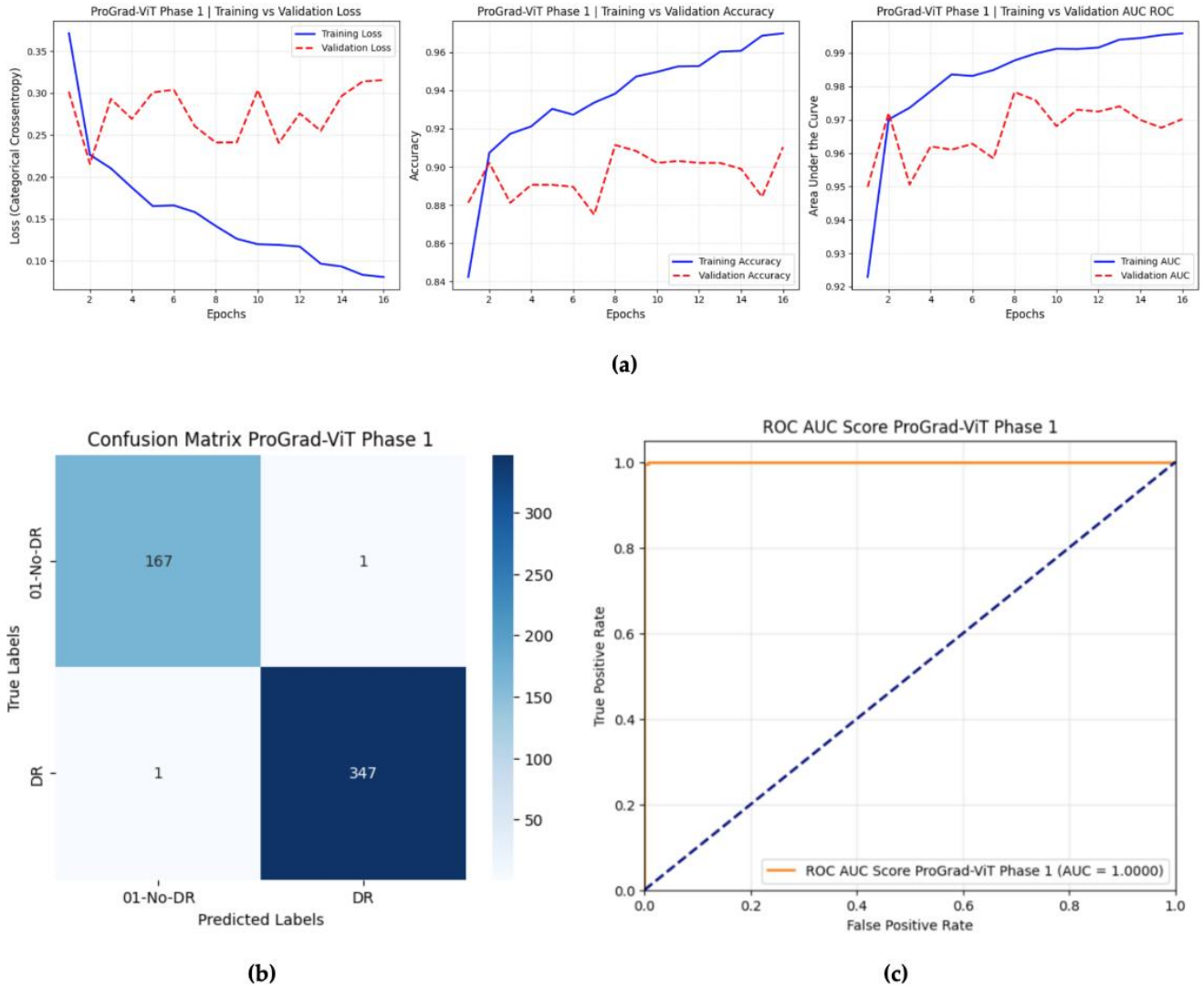

Figure 3. *Visual evaluation of the Phase 1 binary screening model (InceptionV3 + ViT) on the IDRiD hold-out test dataset. (a) Training and validation learning curves for categorical cross-entropy loss and accuracy, demonstrating stable model convergence. (b) Confusion matrix illustrating the high diagnostic agreement between true and predicted labels for healthy (01-No-DR) and pathological (DR) cases. (c) Receiver Operating Characteristic (ROC) curve showing an Area Under the Curve (AUC) score near 1.0000.*

4.1.2. Phase 2: ProGrad-ViT Multi-class Grading Evaluation

Following the robust knowledge initialisation achieved in Phase 1, the ProGrad-ViT architecture was fine-tuned as a single, unified hybrid network to directly classify all five categories of Diabetic Retinopathy, bypassing the need for complex hierarchical sub-network structures. This phase evaluated the model's capacity to discriminate fine-grained, ordinal pathological features on the strictly isolated IDRiD hold-out test set.

The detailed class-wise clinical metrics are presented in Table 4. Despite the inherent clinical complexities of distinguishing intermediate severity stages, the model achieved a global Balanced Accuracy of 0.7534 across the five classes. The exclusive reliance on algorithmic class weighting to manage the severe native dataset imbalance proved effective, particularly for underrepresented categories like the '02-Mild' class, which secured a Sensitivity (Recall) of 0.8000 and a Specificity of 0.9695 despite comprising only 25 test instances. At the baseline, the '01-No-DR' class maintained exceptional discrimination with a Sensitivity of 0.9940 and a Specificity of 1.0000.

To rigorously evaluate the ordinal grading capability of this direct 5-class classification network, the QWK was assessed. The architecture attained an outstanding QWK score of 0.9048. This metric confirms that the model's inherent penalisation of distant class errors successfully mimics the specialised diagnostic reasoning required for standardised diabetic retinopathy staging.

The uncertainty associated to each estimate is reported next to the corresponding point value in Table 4. The balanced accuracy of 0.7534 has a 95% confidence interval of 0.7073–0.7996, which is strictly disjoint from the interval obtained for the purely convolutional baseline described in Section 4.1.3 (0.5899–0.6926). The interval associated with the 02-Mild sensitivity exhibits the greatest dispersion among the five categories (0.6087–0.9114). This is a direct consequence of the size of that class in IDRiD dataset: the point estimate of 0.8000 corresponds to 20 correctly graded images out of 25, thus a single misclassification would modify the estimate by four percentage points. Consequently, these results should be interpreted as evidence that the progressive learning strategy detects early-stage disease substantially better than the flat baseline, whose CI for the same class (0.2340–0.5926) does not overlap with ours, rather than as a precise estimate of mild-stage sensitivity. The precise magnitude of this gain will require a comprehensive external test cohort with a larger representation of early-stage disease.

Table 4. Clinical evaluation metrics for Phase 2: ProGrad-ViT Multi-class Grading (InceptionV3 + ViT on the IDRiD hold-out test set).

Class	Accuracy	Sensitivity (Recall) [95% CI]	Specificity
01-No-DR	0.9981	0.9940 (0.9778 - 1.0000)	1.0000
02-Mild	0.9612	0.8000 (0.6897 - 0.9630)	0.9695
03-Moderate	0.8372	0.6667 (0.5957 - 0.7337)	0.9195
04-Severe	0.8469	0.7097 (0.6196 - 0.8056)	0.8771
05-Proliferative	0.9147	0.5968 (0.4736 - 0.7174)	0.9581
Global Metrics Balanced Accuracy: 0.7534 (0.7141 - 0.8060) Macro-average ROC-AUC: 0.9389 (0.9244 - 0.9521) QWK: 0.9048 (0.8841 - 0.9305)			

A comprehensive visual analysis of the multi-class grading performance is depicted in Figure 4. The multi-class learning curves (Figure 4a) illustrate the continuous categorical cross-entropy optimisation and accuracy stabilisation over the 30-epoch fine-tuning regime. The confusion matrix (Figure 4b) reveals a crucial clinical pattern: misclassifications primarily occur between adjacent ordinal severity levels (e.g., Moderate and Severe), which aligns with the inter-grader variance commonly observed among human specialists. Notably, the model successfully avoids catastrophic diagnostic errors, ensuring no severe or proliferative cases are completely misclassified as healthy retinas.

The One-vs-Rest ROC curves (Figure 4c) underscore the strong discriminative capability of the spatial and sequential features extracted by the hybrid architecture, yielding optimal ROC-AUC scores of 0.9985 for No-DR and 0.9836 for Mild DR. Furthermore, the more complex pathological stages, Severe and Proliferative, achieved robust ROC-AUC of 0.9034 and 0.9025, respectively. Overall, the multi-class evaluation yielded a macro-average ROC-AUC of 0.9389, demonstrating consistent discriminative power across all severity levels.

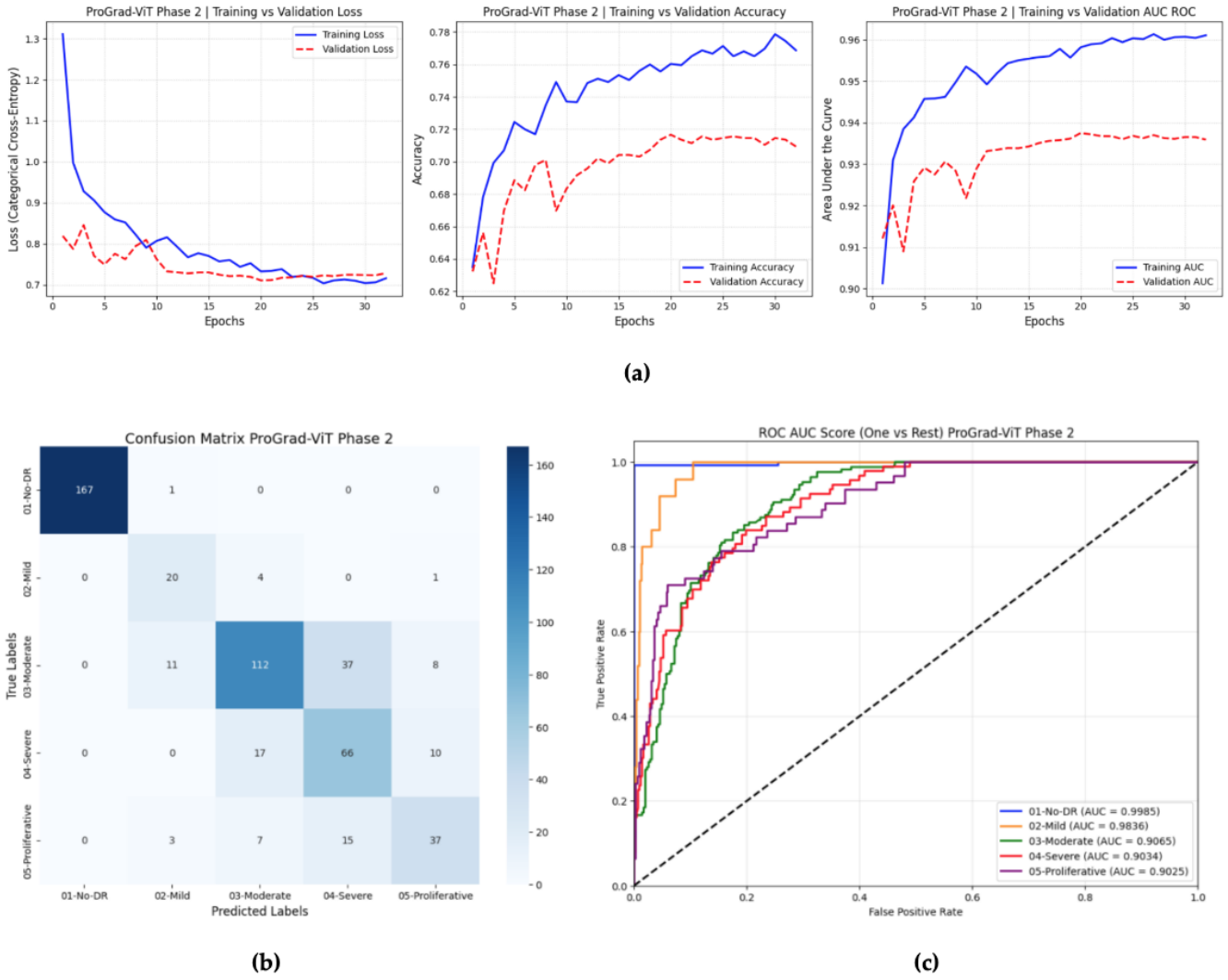

Figure 4. *Visual evaluation of the Phase 2 multi-class grading model (ProGrad-ViT) on the IDRiD hold-out test dataset. (a) Training and validation learning curves over 30 epochs, depicting the optimisation of categorical cross-entropy loss and accuracy stabilisation. (b) Multi-class confusion matrix revealing that misclassifications predominantly occur between adjacent ordinal severity levels. (c) One-vs-Rest Receiver Operating Characteristic (ROC) curves, demonstrating high Area Under the Curve (AUC) scores across all five diabetic retinopathy stages.*

4.1.3. Ablation Study

To empirically validate the architectural necessity of integrating both the ViT and the Progressive Transfer Learning regime, two ablation experiments were conducted. First, a baseline flat CNN utilising only the InceptionV3 backbone without the self-attention encoder was evaluated. Second, a Direct Hybrid model (InceptionV3 + ViT) was trained directly on the 5-class ordinal grading task from ImageNet initialisation, explicitly skipping the Phase 1 binary screening triage. Both configurations employed the same algorithmic class-weighting strategy and were evaluated on the independent IDRiD hold-out test set to ensure a fair comparison.

The class-wise clinical metrics for the baseline model are detailed in Table 5. While the pure CNN architecture demonstrated adequate performance in identifying healthy retinas (01-No-DR) with a Sensitivity of 1.0000, its diagnostic capability severely degraded across the intermediate and minority pathological stages. The absence of the ViT's global spatial reasoning resulted in a significant drop in the global Balanced Accuracy, falling to 0.6413 compared to the 0.7534 achieved by the ProGrad-ViT hybrid framework.

This degradation is particularly critical in the '02-Mild' and '03-Moderate' stages. The baseline model achieved a Sensitivity (Recall) of only 0.4000 and 0.5000 for these classes, respectively. This indicates a high rate of false negatives where the flat CNN, constrained by its localised receptive field, failed to correlate scattered, subtle microaneurysms, misclassifying early-stage pathology as healthy.

Table 5. Clinical evaluation metrics for the Ablation Studies (Baseline Flat InceptionV3 vs. Direct Hybrid InceptionV3+ViT on the IDRiD hold-out test set).

Baseline Flat InceptionV3			
Class	**Accuracy**	**Sensitivity (Recall) [95% CI]**	**Specificity**
01-No-DR	0.9729	1.0000 (1.0000 - 1.0000)	0.9598
02-Mild	0.9554	0.4000 (0.2105 - 0.6000)	0.9837
03-Moderate	0.8023	0.5000 (0.4286 - 0.5834)	0.9483
04-Severe	0.8062	0.6452 (0.5544 - 0.7447)	0.8416
05-Proliferative	0.8702	0.6613 (0.5416 - 0.7714)	0.8987
Global Metrics Balanced Accuracy: 0.6413 (0.5938 - 0.6967) Macro-average ROC-AUC: 0.9058 (0.8961 - 0.9322) QWK: 0.8866 (0.8620 - 0.9064)			
Direct Hybrid InceptionV3+ViT			
Class	**Accuracy**	**Sensitivity (Recall) [95% CI]**	**Specificity**
01-No-DR	0.9748	0.9940 (0.9794 - 1.0000)	0.9655
02-Mild	0.9457	0.4400 (0.2500 - 0.6316)	0.9715
03-Moderate	0.7074	0.1369 (0.0864 - 0.1914)	0.9828
04-Severe	0.7597	0.6667 (0.5714 - 0.7647)	0.7801
05-Proliferative	0.8023	0.7097 (0.6000 - 0.8125)	0.8150
Global Metrics Balanced Accuracy: 0.5895 (0.5379 - 0.6397) Macro-average ROC-AUC: 0.8985 (0.8778 - 0.9180) QWK: 0.8359 (0.8074 - 0.8636)			

The visual evaluation presented in Figure 5 further highlights the structural limitations of the baseline CNN. While the learning curves (Figure 5a) depict optimisation and convergence behaviour, this fails to translate into diagnostic accuracy for intermediate stages. The confusion matrix (Figure 5b) illustrates a pronounced dispersion of predictions, proving that the flat CNN, restricted by its localised receptive field, struggles to establish clear ordinal boundaries. Consequently, the baseline architecture achieved a Macro-Average ROC-AUC of 0.9058 (Figure 5c) and a QWK of 0.8866. While representing a strong agreement, this reduction from the progressive hybrid model demonstrates that the flat CNN incurs in more severe, distant-class misclassifications.

Furthermore, the second ablation configuration, i.e., the Direct Hybrid model, corroborated the methodological necessity of the progressive regime. Without the Phase 1 pre-training to stabilise the self-attention mechanisms, the direct hybrid network suffered from severe instability and a loss of discriminative capacity for the minority clinical stages. While it retained a high sensitivity for the majority '01-No-DR' class (0.9940), its diagnostic capacity for intermediate stages declined substantially. Specifically, sensitivity for the '03-Moderate' class decreased to 0.1369, while the clinically subtle '02-Mild' class dropped to 0.4400. Consequently, the global metrics degraded significantly: Balanced Accuracy fell to 0.5895, Macro-Average ROC-AUC dropped to 0.8985, and the Quadratic Weighted Kappa (QWK) degraded to 0.8359. These findings demonstrate that simply appending a Vision Transformer to a CNN backbone is insufficient for highly imbalanced medical datasets and that the sequential progressive strategy is strictly essential to harness the ViT global spatial reasoning effectively.

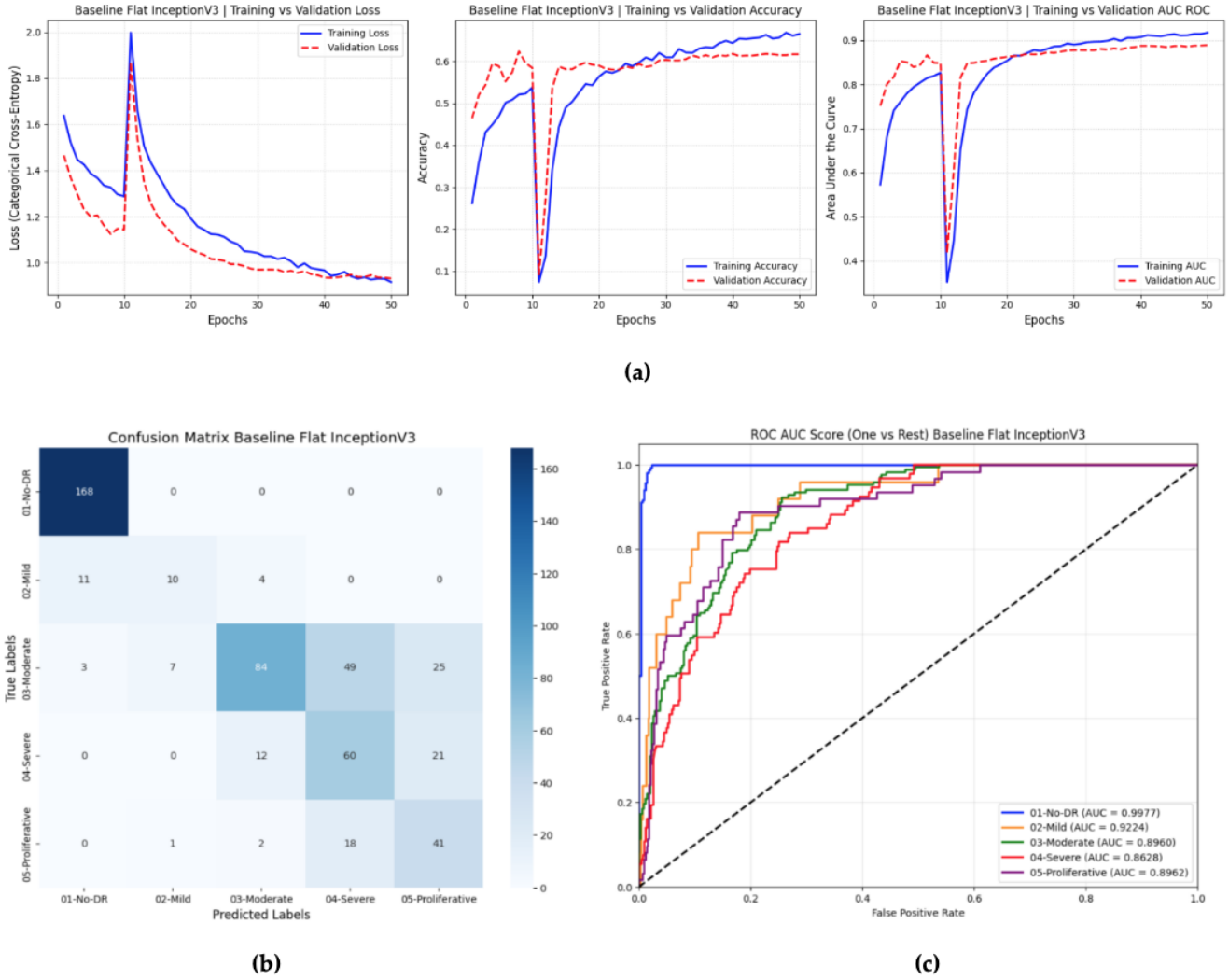

Figure 5. *Visual evaluation of the baseline flat Convolutional Neural Network (InceptionV3 without Vision Transformer) on the IDRiD hold-out test dataset. (a) Training and validation learning curves over 50 epochs, illustrating optimisation behaviour. (b) Multi-class confusion matrix showing elevated misclassifications, particularly in underrepresented intermediate severity stages (02-Mild and 03-Moderate). (c) One-vs-Rest Receiver Operating Characteristic (ROC) curves, displaying reduced Area Under the Curve (AUC) performance compared to the proposed hybrid architecture.*

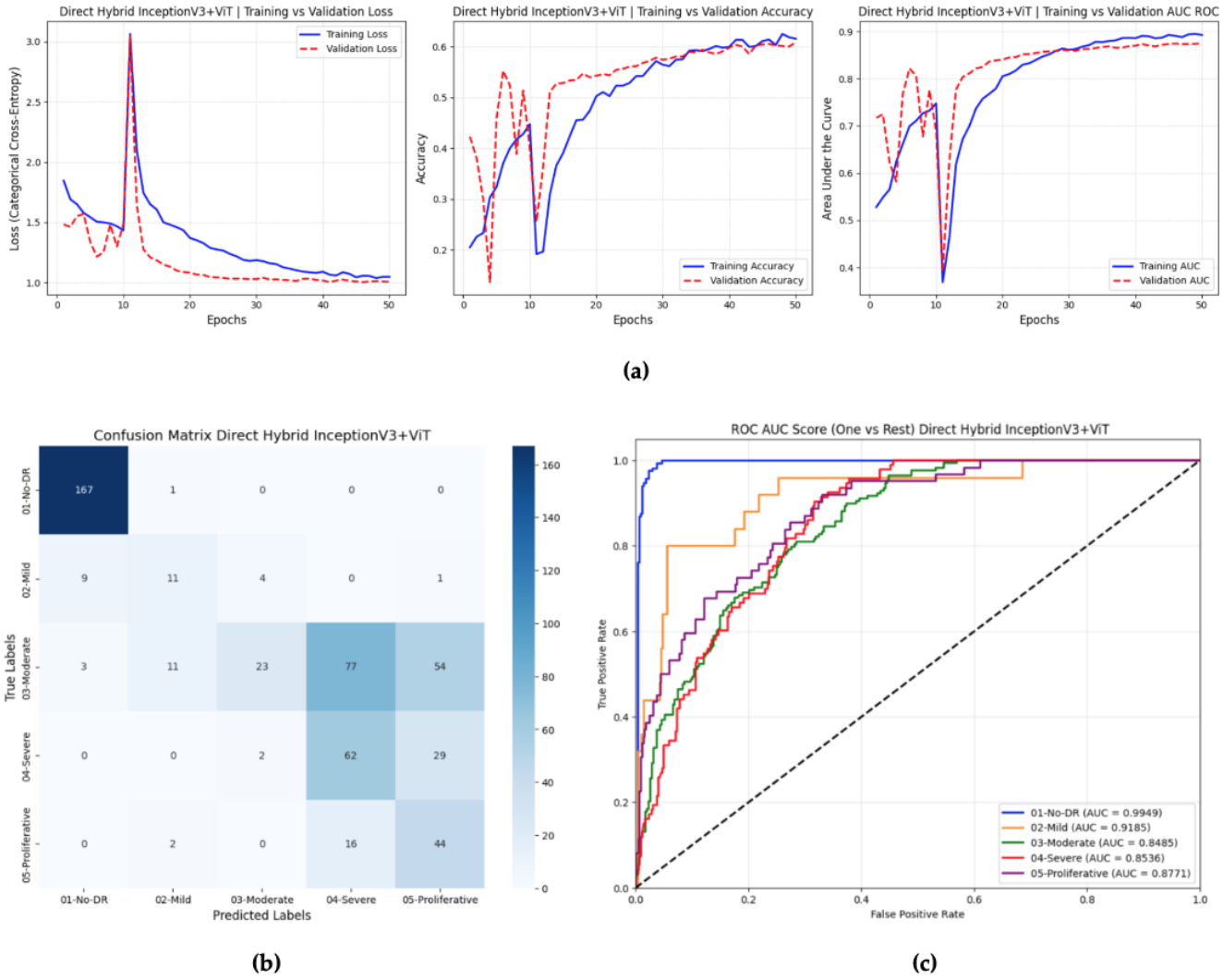

Figure 6. *Performance evaluation of the Direct Hybrid (InceptionV3 + ViT) ablation model. (a) Training and validation learning curves (Loss, Accuracy, and AUC) demonstrating early instability and overfitting in the absence of progressive learning. (b) Confusion matrix on the independent IDRiD hold-out test set illustrating the collapse of predictions for intermediate minority classes (02-Mild and 03-Moderate). (c) One-vs-Rest Receiver Operating Characteristic (ROC) curves, yielding a macro-average AUC of 0.8985.*

Table 6. Isolated contribution of the Vision Transformer encoder and of the Progressive Transfer Learning regime, evaluated on the IDRiD hold-out test set.

Configuration	**Balanced Accuracy [95% CI]**	**Macro-average ROC-AUC [95% CI]**	**QWK [95% CI]**	**02-Mild Sensitivity [95% CI]**
(i) InceptionV3 only (no ViT, no progressive transfer)	0.6413 (0.5938 - 0.6967)	0.9058 (0.8961 - 0.9322)	0.8866 (0.8620 - 0.9064)	0.4000 (0.2105 - 0.6000)
(ii) InceptionV3 + ViT, trained directly on the 5-class task	0.5895 (0.5379 - 0.6397)	0.8985 (0.8778 - 0.9180)	0.8359 (0.8074 - 0.8636)	0.4400 (0.2500 - 0.6316)
(iii) ProGrad-ViT (InceptionV3 + ViT + progressive transfer)	0.7534 (0.7141 - 0.8060)	0.9389 (0.9244 - 0.9521)	0.9048 (0.8841 - 0.9305)	0.8000 (0.6897 - 0.9630)

Removing the progressive training stage while retaining the Transformer encoder yields a balanced accuracy of 0.5895, a macro-average ROC-AUC of 0.8985 and a QWK of 0.8359 on the same hold-out set. The three rows of Table 6 therefore separate the two distinct effects: the performance gap between configurations (i) and (ii) is attributable to the self-attention

encoder alone, whereas the gap between (ii) and (iii) isolates the contribution of the progressive strategy. This effect is most visible in the '02-Mild' class, where sensitivity increases from 0.4000 in the flat CNN to 0.4400 in the directly trained hybrid, finally reaching 0.8000 in the progressive method. This is consistent with our interpretation that the binary pre-training stabilises the attention maps before the network is required to resolve fine ordinal boundaries, rather than simply increasing capacity.

4.2. Discussion

4.2.1. Architectural Paradigm and Feature Integration

The accurate grading of Diabetic Retinopathy (DR) requires a comprehensive evaluation of both localised retinal lesions and their widespread distribution. While traditional Convolutional Neural Networks (CNNs) excel at extracting fine-grained local features, they inherently suffer from a limited receptive field, restricting their ability to capture long-range dependencies across the global spatial context of the retina. The proposed ProGrad-ViT addresses this critical limitation by unifying an InceptionV3 backbone with a Vision Transformer (ViT) encoder in a single-stage, 5-class classification network. The results demonstrate that this hybrid architecture effectively leverages the strong inductive biases of the CNN to detect isolated anomalies, such as microaneurysms and haemorrhages, while the ViT utilises self-attention mechanisms to dynamically correlate scattered lesions across different retinal quadrants.

4.2.2. Progressive Learning and Data Integrity

A major challenge in applying ViTs to medical imaging is their "data-hungry" nature, which frequently leads to severe overfitting when trained from scratch on small or highly imbalanced datasets. To overcome this, our methodology implemented a Progressive Transfer Learning regime. By pre-training the network on a binary screening task, the Transformer's attention mechanisms were stabilised prior to the more complex ordinal grading task. A second methodological contribution concerns the treatment of class imbalance. The extreme morphological variation between DR severity stages has historically pushed state-of-the-art approaches towards synthetic data manipulation, such as physical oversampling or targeted augmentation. These practices distort the natural morphological distribution of retinal biomarkers. That distortion compromises the reliability of the resulting model in real-world scenarios. In contrast, ProGrad-ViT utilises strictly algorithmic class weighting to penalise misclassifications in minority classes. This strategy ensures that the morphology of the original images is preserved, yielding clinical metrics that reflect the true diagnostic capability of the architecture.

4.2.3. Comparative Analysis with State-of-the-Art

The superiority of this unified approach is evidenced when comparing ProGrad-ViT against recent state-of-the-art models, as can be seen in Table 7.

Table 7. Performance comparison of ProGrad-ViT against state-of-the-art DR grading models.

Reference	Architecture Paradigm	Imbalance Strategy	Dataset (Test)	Accuracy	Balanced Accuracy	QWK
Ashwini & Dash (2023)	Pure CNNs (ResNet, DenseNet, etc.)	Not specified	IDRiD	0.9007	N/R	N/R
Yi et al. (2026)	Hybrid (CTANet: CNN + ViT)	Not specified	APTOS 2019	0.9008	N/R	N/R
Ikram & Imran (2025)	Hybrid (ResViT: ResNet50 + ViT)	Physical Data Augmentation	APTOS 2019	0.9301	N/R	0.8935
Girija et al. (2025)	CNN + Attention (EfficientNetB5-CBAM)	Targeted Augmentation	Not specified	0.7069	N/R	0.8451
Liu et al. (2025)	Swin-Transformer V2	Not specified	Multiple (incl. APTOS)	0.7700	N/R	0.8770
Ours	**Hybrid (InceptionV3 + ViT + Progressive regime)**	**Algorithmic Class Weights**	**IDRiD**	**0.7810**	**0.7534**	**0.9048**

N/R: not reported in the corresponding publication. All figures are those stated by the original authors on the test set indicated in the fourth column.

As demonstrated in Table 7, enhancing high-capacity CNNs with purely convolutional attention mechanisms (EfficientNetB5-CBAM) alongside targeted augmentation (Girija et al., 2025) achieved a QWK of 0.8451. Similarly, pure Transformer architectures like STMF-DRNet (Liu et al., 2025) yielded a QWK of 0.8770. Notably, the ResViT FusionNet (Ikram & Imran, 2025), a hybrid model that applied physical data augmentation to balance the dataset, reached a QWK of 0.8935. ProGrad-ViT outperforms these methodologies by achieving a QWK of 0.9048. This metric is particularly significant for clinical triage, as QWK heavily penalises severe misclassifications between distant ordinal grades. The results confirm that the proposed network not only matches but exceeds the ordinal separation capabilities of heavily augmented models, validating the efficacy of our unified architecture and purely algorithmic weighting strategy.

4.2.4. Limitations and Future Directions

While the proposed framework demonstrates robust performance, the study presents certain limitations, the first of which concerns the nature of the evidence provided. The evaluation reported here establishes cross-dataset technical generalisation: the model was trained on APTOS 2019 and MESSIDOR and evaluated, without any adaptation, on IDRiD, a dataset acquired in a different country, with different cameras and by a different grading team. This is a demanding form of retrospective validation and a necessary condition for clinical usefulness, but it is not equivalent to clinical validation. Public benchmarks are curated collections in which images of insufficient quality are typically discarded before release, the severity distribution does not reproduce that of a screening programme, and the reference labels come from a small number of graders without a formal adjudication protocol. Demonstrating that the framework is safe for triage would require a prospective study on consecutive, unselected patients, with pre-registered endpoints, an explicit analysis of the behaviour on ungradable images, and a comparison against the ophthalmologists who currently perform the task in the target setting. The conclusions drawn in this study should be interpreted accordingly: ProGrad-ViT is a candidate for such a study, rather than a validated clinical tool.

A second limitation concerns the size and composition of the test set. IDRiD contains 516 images, and only 25 of them belong to the 02-Mild class, which is precisely the category with the greatest clinical value for early referral. The confidence intervals reported in Table 4 make the consequence explicit: the point estimates for the minority stages are informative about the ordering of the competing configurations, but not precise enough to support a specific operating threshold. External evaluation on larger and more balanced cohorts, ideally from screening programmes in several regions, is required before the class-wise metrics can be treated as stable.

A third limitation is architectural. The convolutional backbone was fixed to InceptionV3 for the reasons set out in Section 3.3.1, and the contribution of the progressive regime was measured with that base architecture held constant. Whether the same gains persist with EfficientNetV2, ConvNeXt or a Swin-V2 encoder remains an open question, and a systematic benchmark of alternative backbones under an identical progressive protocol is the most immediate extension of this work.

Finally, the algorithmic weighting strategy, while preserving morphological integrity, relies heavily on the optimisation landscape, which can still present convergence challenges for extreme minority classes. Future work will focus on integrating multi-modal data streams to provide a more comprehensive clinical assessment and exploring computationally lighter ViT variants to facilitate deployment in resource-constrained telemedicine environments.

5 Conclusions

In this study, we presented ProGrad-ViT, a novel hybrid deep learning architecture designed for the automated and accurate grading of Diabetic Retinopathy (DR). By unifying an InceptionV3 backbone with a Vision Transformer encoder within a single-stage, 5-class classification framework, the proposed model successfully bridges the gap between local lesion detection and global spatial contextualisation. This architectural synergy allows for a comprehensive evaluation of the retina, capturing both isolated anomalies and their widespread distribution patterns, which is critical for distinguishing between closely related advanced DR stages.

A fundamental contribution of our methodology is the strategic approach to extreme class imbalance and data scarcity. Rather than relying on synthetic data manipulation or physical oversampling techniques that can artificially distort the pristine morphological distribution of retinal biomarkers, this work implemented a strictly algorithmic class-weighting strategy. Coupled with a Progressive Transfer Learning regime, transitioning from binary screening to complex ordinal grading, this approach stabilised the Transformer's attention mechanisms and ensured that the model learned from the true clinical representation of the disease.

The experimental results validate the efficacy of this paradigm. Evaluated on the IDRiD dataset, ProGrad-ViT achieved a global Accuracy of 0.7810, a Balanced Accuracy of 0.7534 (95% CI 0.7073–0.7996), and a highly competitive QWK of 0.9048. These metrics demonstrate that the model not only effectively mitigates the bias towards majority classes but also provides exceptional ordinal separation capabilities, outperforming heavily augmented state-of-the-art models.

Ultimately, ProGrad-ViT offers a robust and reproducible basis for automated DR triage under a demanding cross-dataset protocol. Confirming its clinical value will require prospective evaluation on unselected patients, as discussed earlier. Future research will explore the integration of multi-modal data streams and the optimisation of computationally lighter Transformer variants to facilitate real-time deployment in resource-constrained telemedicine environments.

References

Abràmoff, M. D., Garvin, M. K., & Sonka, M. (2010). Retinal imaging and image analysis. *IEEE Reviews in Biomedical Engineering, 3*, 169–208. https://doi.org/10.1109/RBME.2010.2084567

Karthik, Maggie, & Dane, S. (2019). *APTOS 2019 blindness detection* [Data set]. Kaggle. https://www.kaggle.com/competitions/aptos2019-blindness-detection

Ashwini, K., & Dash, R. (2023). Grading diabetic retinopathy using multiresolution based CNN. *Biomedical Signal Processing and Control, 86*, Article 105210. https://doi.org/10.1016/j.bspc.2023.105210

Azad, R., Kazerouni, A., Heidari, M., Aghdam, E. K., Molaei, A., Jia, Y., Jose, A., Roy, R., & Merhof, D. (2024). Advances in medical image analysis with vision transformers: A comprehensive review. *Medical Image Analysis, 91*, Article 103000. https://doi.org/10.1016/j.media.2023.103000

Banerjee, T., Singh, D. P., & Kour, P. (2026). Advances in deep neural, transformer learning, and kernel-based methods for diabetic retinopathy detection: A comprehensive review. *Archives of Computational Methods in Engineering, 33*, 2659–2707. https://doi.org/10.1007/s11831-025-10376-8

Bhardwaj, C., Jain, S., & Sood, M. (2021). Deep learning–based diabetic retinopathy severity grading system employing quadrant ensemble model. *Journal of Digital Imaging, 34*(2), 440–457. https://doi.org/10.1007/s10278-021-00418-5

Bhoyar, V., & Patel, M. (2026). A comprehensive review of deep learning approaches for automated detection, segmentation, and grading of diabetic retinopathy. *Archives of Computational Methods in Engineering, 33*, 4359–4380. https://doi.org/10.1007/s11831-025-10460-z

Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., & Malanda, B. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Research and Clinical Practice, 138*, 271–281. https://doi.org/10.1016/j.diabres.2018.02.023

Chowdhury, A. R., Chatterjee, T., & Banerjee, S. (2019). A random forest classifier-based approach in the detection of abnormalities in the retina. *Medical & Biological Engineering & Computing, 57*(1), 193–203. https://doi.org/10.1007/s11517-018-1878-0

Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., Charton, B., & Klein, J.-C. (2014). Feedback on a publicly distributed image database: The Messidor database. *Image Analysis & Stereology, 33*(3), 231–234. https://doi.org/10.5566/ias.1155

Fong, D. S., Aiello, L., Gardner, T. W., King, G. L., Blankenship, G., Cavallerano, J. D., Ferris, F. L., III, Klein, R., & American Diabetes Association. (2004). Retinopathy in diabetes. *Diabetes Care, 27*(Suppl. 1), S84–S87. https://doi.org/10.2337/diacare.27.2007.s84

Girija, R., Deepa, N., & Chinni, S. K. (2025). CBAM-integrated deep neural networks for diabetic retinopathy detection with quadratic weighted kappa evaluation. In *2025 3rd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)* (pp. 304–309). IEEE. https://doi.org/10.1109/ICAICCIT68829.2025.11433973

Guariguata, L., Whiting, D., Weil, C., & Unwin, N. (2011). The International Diabetes Federation diabetes atlas methodology for estimating global and national prevalence of diabetes in adults. *Diabetes Research and Clinical Practice, 94*(3), 322–332. https://doi.org/10.1016/j.diabres.2011.10.040

Hagos, M. T., & Kant, S. (2019). *Transfer learning based detection of diabetic retinopathy from small dataset* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1905.07203

He, A., Li, T., Li, N., Wang, K., & Fu, H. (2021). CABNet: Category attention block for imbalanced diabetic retinopathy grading. *IEEE Transactions on Medical Imaging, 40*(1), 143–153. https://doi.org/10.1109/TMI.2020.3023463

Ikram, A., & Imran, A. (2025). ResViT FusionNet model: An explainable AI-driven approach for automated grading of diabetic retinopathy in retinal images. *Computers in Biology and Medicine, 186*, Article 109656. https://doi.org/10.1016/j.compbiomed.2025.109656

International Diabetes Federation. (2013). Five questions on the IDF Diabetes Atlas. *Diabetes Research and Clinical Practice, 102*(2), 147–148. https://doi.org/10.1016/j.diabres.2013.10.013

International Diabetes Federation. (2025). *IDF Diabetes Atlas* (11th ed.). https://diabetesatlas.org/resources/idf-diabetes-atlas-2025/

Khanna, M., Singh, L. K., Thawkar, S., & Goyal, M. (2023). Deep learning based computer-aided automatic prediction and grading system for diabetic retinopathy. *Multimedia Tools and Applications, 82*, 39255–39302. https://doi.org/10.1007/s11042-023-14970-5

Kwan, C. C., & Fawzi, A. A. (2019). Imaging and biomarkers in diabetic macular edema and diabetic retinopathy. *Current Diabetes Reports, 19*(10), Article 95. https://doi.org/10.1007/s11892-019-1226-2

Lechner, J., O'Leary, O. E., & Stitt, A. W. (2017). The pathology associated with diabetic retinopathy. *Vision Research, 139*, 7–14. https://doi.org/10.1016/j.visres.2017.04.003

Li, Y.-H., Yeh, N.-N., Chen, S.-J., & Chung, Y.-C. (2019). Computer-assisted diagnosis for diabetic retinopathy based on fundus images using deep convolutional neural network. *Mobile Information Systems, 2019*, Article 6142839. https://doi.org/10.1155/2019/6142839

Liu, T., Chen, Y., Shen, H., Zhou, R., Zhang, M., Liu, T., & Liu, J. (2021). A novel diabetic retinopathy detection approach based on deep symmetric convolutional neural network. *IEEE Access, 9*, 160552–160558. https://doi.org/10.1109/ACCESS.2021.3131630

Liu, Y., Yao, D., Ma, Y., Wang, H., Wang, J., Bai, X., Zeng, G., & Liu, Y. (2025). STMF-DRNet: A multi-branch fine-grained classification model for diabetic retinopathy using Swin-TransformerV2. *Biomedical Signal Processing and Control, 103*, Article 107352. https://doi.org/10.1016/j.bspc.2024.107352

Majumder, S., & Kehtarnavaz, N. (2021). Multitasking deep learning model for detection of five stages of diabetic retinopathy. *IEEE Access, 9*, 123220–123230. https://doi.org/10.1109/ACCESS.2021.3109240

Mansour, R. F. (2018). Deep-learning-based automatic computer-aided diagnosis system for diabetic retinopathy. *Biomedical Engineering Letters, 8*, 41–57. https://doi.org/10.1007/s13534-017-0047-y

Mazlan, N., Yazid, H., Arof, H., & Isa, H. (2020). Automated microaneurysms detection and classification using multilevel thresholding and multilayer perceptron. *Journal of Medical and Biological Engineering, 40*(2), 292–306. https://doi.org/10.1007/s40846-020-00509-8

Nazih, W., Aseeri, A. O., Atallah, O. Y., & El-Sappagh, S. (2023). Vision transformer model for predicting the severity of diabetic retinopathy in fundus photography-based retina images. *IEEE Access, 11*, 117546–117561. https://doi.org/10.1109/ACCESS.2023.3326528

Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., & Meriaudeau, F. (2018). Indian Diabetic Retinopathy Image Dataset (IDRiD): A database for diabetic retinopathy screening research. *Data, 3*(3), Article 25. https://doi.org/10.3390/data3030025

Reguant, R., Brunak, S., & Saha, S. (2021). Understanding inherent image features in CNN-based assessment of diabetic retinopathy. *Scientific Reports, 11*, Article 9704. https://doi.org/10.1038/s41598-021-89225-0

Sambyal, N., Saini, P., Syal, R., & Gupta, V. (2020). Modified U-Net architecture for semantic segmentation of diabetic retinopathy images. *Biocybernetics and Biomedical Engineering, 40*(3), 1094–1109. https://doi.org/10.1016/j.bbe.2020.05.006

Seoud, L., Chelbi, J., & Cheriet, F. (2015). Automatic grading of diabetic retinopathy on a public database. In X. Chen, M. K. Garvin, J. J. Liu, E. Trucco, & Y. Xu (Eds.), *Proceedings of the Ophthalmic Medical Image Analysis Second International Workshop (OMIA 2015)* (pp. 97–104). https://doi.org/10.17077/omia.1032

Sinclair, A., Saeedi, P., Kaundal, A., Karuranga, S., Malanda, B., & Williams, R. (2020). Diabetes and global ageing among 65–99-year-old adults: Findings from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Research and Clinical Practice, 162*, Article 108078. https://doi.org/10.1016/j.diabres.2020.108078

Skouta, A., Elmoufidi, A., Jai-Andaloussi, S., & Ouchetto, O. (2022). Hemorrhage semantic segmentation in fundus images for the diagnosis of diabetic retinopathy by using a convolutional neural network. *Journal of Big Data, 9*, Article 78. https://doi.org/10.1186/s40537-022-00632-0

Usman, T. M., Saheed, Y. K., Ignace, D., & Nsang, A. (2023). Diabetic retinopathy detection using principal component analysis multi-label feature extraction and classification. *International Journal of Cognitive Computing in Engineering, 4*, 78–88. https://doi.org/10.1016/j.ijcce.2023.02.002

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.

Wang, Y.-B., Zhu, C.-Z., Yan, Q.-F., & Liu, L.-Q. (2016). A novel vessel segmentation in fundus images based on SVM. In *2016 International Conference on Information System and Artificial Intelligence (ISAI)* (pp. 390–394). IEEE. https://doi.org/10.1109/ISAI.2016.0089

Watkins, P. J. (2003). Retinopathy. *BMJ, 326*(7395), 924–926. https://doi.org/10.1136/bmj.326.7395.924

Wilkinson, C. P., Ferris, F. L., III, Klein, R. E., Lee, P. P., Agardh, C. D., Davis, M., Dills, D., Kampik, A., Pararajasegaram, R., Verdaguer, J. T., & Global Diabetic Retinopathy Project Group. (2003). Proposed international clinical

diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology, 110*(9), 1677–1682. https://doi.org/10.1016/S0161-6420(03)00475-5

Yi, S., Ren, Y., & Shao, D. (2026). CTANet: A hybrid CNN and ViT network with adaptive focus fusion for diabetic retinopathy grading. *Biomedical Signal Processing and Control, 113*, Article 109106. https://doi.org/10.1016/j.bspc.2025.109106