Comparative Study of Deep Learning Architectures for Automated Diabetic Retinopathy Grading: Vision Transformer, Swin Transformer, and InceptionResNetV2

Mahamadtohid Naikwadi
Department of Computer Engineering
KIT's College of Engineering
Kolhapur, Maharashtra, India
naikwadimdtohid@gmail.com

Prajwal Khandait
Department of Computer Engineering
KIT's College of Engineering
Kolhapur, Maharashtra, India
khandaitprajwal@gmail.com

Aditya Sutar
Department of Computer Engineering
KIT's College of Engineering
Kolhapur, Maharashtra, India
adityasutaraiml77@gmailcom

Samir Mulla
Department of Computer Engineering
KIT's College of Engineering
Kolhapur, Maharashtra, India
mullasamir799@gmail.com

Rajesh Kumar
Department of Computer Engineering
KIT's College of Engineering
Kolhapur, Maharashtra, India
kumar.rajesh@kitcoek.in

Uma Gurav
Department of Computer Engineering
KIT's College of Engineering
Kolhapur, Maharashtra, India
gurav.uma@kitcoek.in

***Abstract*—Diabetic Retinopathy (DR) is a vision-threatening complication of diabetes mellitus that progresses silently through five clinically defined severity grades. Timely automated screening is critical to prevent irreversible vision loss, particularly in resource-constrained healthcare settings. This paper presents a systematic comparative study of three state-of-the-art deep learning architectures-Vision Transformer (ViT-Base/16), Swin Transformer (swin_base_patch4_window7_224), and InceptionResNetV2-applied to five-class DR grading on the APTOS 2019 fundus image dataset (3,662 images). All models employ transfer learning from ImageNet-pretrained weights. We analyze each architecture from the perspectives of classification accuracy, per-class F1-score, macro-averaged AUC, GradCAM-based explainability, training dynamics, and parameter efficiency. Our ViT-Base/16 model, fine-tuned end-to-end with AdamW, cosine annealing, and label smoothing, achieves the highest validation accuracy of 85.40% with a macro-averaged F1-score of 0.7247. Swin Transformer achieves 83.20% accuracy, while InceptionResNetV2 achieves 81.40% through two-stage transfer learning. GradCAM visualizations confirm clinically aligned lesion localization across all architectures. This work provides architectural insights for deploying robust DR screening systems in clinical environments.**

***Index Terms*—Diabetic Retinopathy, Vision Transformer, Swin Transformer, InceptionResNetV2, Transfer Learning, GradCAM, Fundus Image Classification, Deep Learning, Medical Image Analysis**

I. Introduction

Diabetes mellitus is a globally prevalent metabolic disorder affecting over 537 million adults worldwide, with projections suggesting this number will reach 783 million by 2045 [1]. Diabetic Retinopathy (DR), a microvascular complication of diabetes, is the leading cause of preventable blindness among working-age adults [2]. DR develops progressively through microaneurysms, hemorrhages, hard exudates, neovascularization and ultimately vitreous hemorrhage or tractional retinal detachment [3]. The International Clinical DR Severity Scale (ICDRSS) classifies the disease into five grades: No DR, Mild Non-Proliferative DR (NPDR), Moderate NPDR, Severe NPDR, and Proliferative DR (PDR). Early detection at Mild or Moderate stages enables timely laser photocoagulation or anti-VEGF therapy, significantly reducing the risk of vision loss [4].

Manual grading of fundus photographs by trained ophthalmologists is expensive, time-consuming, and subject to inter-grader variability of approximately 20% at borderline grade boundaries [5]. Automated computer-aided DR detection offers a scalable solution for mass screening, especially in low-income regions where ophthalmologist density is critically low.

Deep learning has transformed medical image analysis, beginning with Convolutional Neural Networks (CNNs) such as VGGNet, ResNet, and Inception families, which achieved dermatologist-level performance on various diagnostic tasks [6]. However, CNNs capture features hierarchically through local receptive fields, limiting their ability to model long-range spatial dependencies between lesions distributed across different retinal quadrants. Vision Transformers (ViTs) [7] address this by treating image patches as tokens and applying global multi-head self-attention, enabling the model to capture retina-wide contextual relationships from the very first layer. Swin Transformer [8] extends this paradigm with a hier-

archical, shifted-window attention mechanism that combines local window efficiency with cross-window context modeling, achieving state-of-the-art results on dense prediction tasks.

This paper makes the following contributions:

- A rigorous comparative study of three architectures-ViT-Base/16, Swin Transformer Base, and InceptionResNetV2-on five-class DR grading using the APTOS 2019 dataset.
- Detailed analysis of training strategies including single-phase end-to-end fine-tuning (ViT) versus two-stage transfer learning (Swin, InceptionResNetV2).
- Per-class F1-score, confusion matrix, and One-vs-Rest AUC evaluation for each model.
- GradCAM/GradCAM++ explainability analysis aligned with clinical DR markers.
- Architectural insights and deployment trade-offs to guide clinical system design.

II. Related Work

A. CNN-Based Approaches

Gulshan et al. [10] demonstrated that a deep CNN trained on 128,175 retinal images achieved an AUC of 0.991 for referable DR detection, matching ophthalmologist-level accuracy in a binary classification setting. Abràmoff et al. [11] deployed IDx-DR, the first FDA-cleared autonomous AI diagnostic system for DR, using a CNN-based pipeline achieving 87.2% sensitivity. Gargeya et al. [12] proposed a custom CNN trained on 75,137 images achieving 94% AUC on the Messidor-2 dataset. These works established that CNNs can reach clinical-grade performance, but are limited to binary (referable/non-referable) classification.

B. Multi-Class Grading

Graham [13] won the Kaggle DR competition using preprocessing-heavy CNNs on the EyePACS dataset (88,702 images). Sikder et al. [14] applied VGG16, ResNet-50, and InceptionV3 on the APTOS 2019 dataset, finding ResNet-50 achieving 82.3% five-class accuracy. Qummar et al. [15] employed an ensemble of five CNNs (DenseNet121, ResNet50, InceptionV3, Xception, InceptionResNetV2) achieving 80.8% on Kaggle DR data. These works established multi-class grading benchmarks that our study extends with transformer-based architectures.

C. Attention and Transformer Approaches

Wang et al. [16] applied self-attention augmented CNNs to fundus images, demonstrating that global attention improves detection of spatially distributed hemorrhages in severe DR. Sun et al. [17] explored ViT on retinal OCT images, finding that transformers generalize better than CNNs when pretrained on large medical datasets. Gheflati et al. [18] applied ViT and Swin Transformer to DR grading on IDRiD and APTOS, finding Swin achieved 85.7% accuracy on APTOS five-class grading. Our work extends this comparison by rigorously evaluating all three architectures under unified dataset splits, augmentation strategies, and explainability analysis.

D. Explainability

Selvaraju et al. [19] proposed GradCAM for CNN spatial localization. Chattopadhay et al. [20] introduced Grad-CAM++, providing more accurate localization for multiple object instances-particularly relevant for DR where multiple lesion types coexist. Chefer et al. [21] developed transformer-specific attribution methods, demonstrating that attention roll-out and gradient-based methods can reliably localize retinal pathologies in ViT.

III. Dataset and Preprocessing

A. Dataset

We use the APTOS 2019 Blindness Detection dataset [24], publicly available on Kaggle (sovitrath/diabetic-retinopathy-224x224-2019-data). The dataset contains **3,662 color fundus photographs** pre-resized to 224×224 pixels with Gaussian spatial filtering applied to enhance retinal vascular contrast. The five-class label distribution is shown in Table I.

TABLE I
APTOS 2019 Dataset Class Distribution

Grade	Class	Images	Share (%)
0	No DR	1,805	49.3%
1	Mild NPDR	370	10.1%
2	Moderate NPDR	999	27.3%
3	Severe NPDR	193	5.3%
4	Proliferative	295	8.1%
	Total	**3,662**	**100%**

A stratified 80/20 train-validation split yields 2,929 training and 733 validation images, preserving class proportions. The dataset exhibits significant class imbalance (No DR: 49.3% vs. Severe: 5.3

B. Preprocessing and Augmentation

All images are standardized to 224×224×3 RGB input. Architecture-specific preprocessing strategies are applied as summarized in Table II.

TABLE II
Preprocessing and Augmentation Strategies per Architecture

Transform	ViT	Swin	IRNetV2
Resize+CenterCrop	256→224	224	224
RandomHorizontalFlip	✓	✓	✓
RandomVerticalFlip	✓(0.3)	-	-
ColorJitter	✓	-	-
RandomRotation	±15°	±10°	±15°
RandomAffine/Shear	✓	-	-
Zoom Range	-	-	±20%
Normalization	ImageNet	ImageNet	$[-1, 1]$
WeightedSampler	✓	-	-
Label Smoothing	0.1	-	-

The ViT pipeline applies the richest augmentation strategy-seven transforms including vertical flip, color jitter,

random affine shear, and a `WeightedRandomSampler` with inverse-frequency class weights to counteract the skewed class distribution. InceptionResNetV2 uses Keras `ImageDataGenerator` with the model-native `preprocess_input` function that scales pixels to $[-1, 1]$. Swin Transformer applies minimal augmentation (flip + rotation) with standard ImageNet normalization.

IV. Proposed Architectures

A. Vision Transformer (ViT-Base/16)

The Vision Transformer [7], specifically `vit_base_patch16_224` from the `timm` library [23], divides each 224×224 input into 196 non-overlapping 16×16 patches. Each patch is linearly projected to a 768-dimensional token embedding, yielding a sequence of 197 tokens after prepending a learnable CLS token. Learnable 1D positional embeddings are added to encode spatial structure.

The encoder consists of 12 transformer blocks, each applying multi-head self-attention (MHSA) with 12 heads (64-dim per head) followed by a 2-layer MLP (768→3072→768, GELU activation), with LayerNorm pre-normalization and residual connections:

$$\mathbf{h}_\ell = \mathbf{h}_{\ell-1} + \text{MHSA}(\text{LN}(\mathbf{h}_{\ell-1})) \tag{1}$$

$$\mathbf{h}'_\ell = \mathbf{h}_\ell + \text{MLP}(\text{LN}(\mathbf{h}_\ell)) \tag{2}$$

The CLS token output from the final encoder block (768-dim) is passed through a custom classification head:

$$\hat{y} = \text{Linear}_{512\to5}(\text{GELU}(\text{Linear}_{768\to512}(\text{Dropout}_{0.30}(\text{LN}(\mathbf{z}_{\text{CLS}}))))) \tag{3}$$

The model's total parameter count is approximately **86.2 million**, pretrained on ImageNet-21K. The key advantage for DR is that every patch attends to every other patch simultaneously across all 12 encoder blocks, enabling direct modeling of long-range spatial relationships between lesions in different retinal quadrants-critical for Severe and Proliferative DR where pathology is spatially distributed.

B. Swin Transformer (swin_base_patch4_window7_224)

The Swin Transformer [8] introduces a hierarchical vision transformer with shifted window (SW-MSA) attention. Unlike ViT's global self-attention over all 196 tokens, Swin partitions image tokens into non-overlapping local windows of 7×7 patches and computes attention within each window, achieving linear complexity $O(n)$ with respect to image size.

The architecture processes images through 4 hierarchical stages separated by patch merging layers that halve spatial resolution while doubling channel dimension, producing multi-scale feature maps analogous to a CNN feature pyramid:

- **Stage 1**: 56×56 patches, 128-dim, 2 blocks (W-MSA + SW-MSA)
- **Stage 2**: 28×28 patches, 256-dim, 2 blocks
- **Stage 3**: 14×14 patches, 512-dim, 18 blocks
- **Stage 4**: 7×7 patches, 1024-dim, 2 blocks

The shifted window mechanism alternates between regular (W-MSA) and shifted (SW-MSA) window configurations across consecutive blocks, enabling cross-window information flow without global computation:

$$\hat{\mathbf{z}}^\ell = \text{W-MSA}(\text{LN}(\mathbf{z}^{\ell-1})) + \mathbf{z}^{\ell-1} \tag{4}$$

$$\hat{\mathbf{z}}^{\ell+1} = \text{SW-MSA}(\text{LN}(\hat{\mathbf{z}}^\ell)) + \hat{\mathbf{z}}^\ell \tag{5}$$

The 5-class linear classification head replaces the original ImageNet head, and the backbone is pretrained on ImageNet-21K. Total parameters: ∼**87 million**. Swin's hierarchical design preserves fine-grained microaneurysm features at early stages while building coarse structural representations at deeper stages.

C. InceptionResNetV2

InceptionResNetV2 [9] is a hybrid deep CNN that fuses multi-scale feature extraction from Inception modules with residual skip connections for stable gradient flow through approximately 164 layers. The model uses factorized convolutions to reduce parameter count while maintaining large receptive fields.

The architecture consists of:

- **Stem**: Initial strided convolutions reducing 224×224 to 35×35
- **Block35 (×5)**: Inception-ResNet-A blocks (multi-branch 1×1, 3×3)
- **ReductionA**: Spatial downsampling 35×35→17×17
- **Block17 (×10)**: Inception-ResNet-B blocks (factorized 1×7, 7×1)
- **ReductionB**: Spatial downsampling 17×17→8×8
- **Block8 (×5)**: Inception-ResNet-C blocks (factorized 1×3, 3×1)
- **Conv7b**: Final 1×1 convolution producing 1536-dim feature maps

The custom classification head consists of GlobalAveragePooling2D followed by Dropout(0.5) and a Dense(5, softmax) output layer. The model is pretrained on ImageNet with `include_top=False`, and the `preprocess_input` function scales inputs to $[-1, 1]$. Total parameters: ∼**54.3 million**.

V. Training Methodology

A. ViT-Base/16: Single-Phase End-to-End Fine-Tuning

All backbone layers are trained simultaneously from epoch 1 using a small learning rate to preserve pretrained representations. A linear warmup schedule ramps the learning rate from 3×10^{-6} to 3×10^{-5} over 3 epochs, followed by cosine annealing decay to $\eta_{\min} = 10^{-6}$ over 12 epochs:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left(1 + \cos\left(\frac{t}{T_{\max}}\pi\right)\right) \tag{6}$$

AdamW optimizer [22] with weight decay $\lambda = 10^{-2}$ and gradient clipping (max_norm = 1.0) prevent attention layer instability. CrossEntropyLoss with label smoothing $\epsilon = 0.1$

addresses ambiguous grade boundaries. Training runs for 15 epochs with early stopping (patience = 10), followed by Phase 2 fine-tuning from the best checkpoint.

B. Swin and InceptionResNetV2: Two-Stage Transfer Learning

Both models follow a two-stage strategy:

Stage 1 - Head Warm-up (5 epochs): The backbone is frozen; only the classification head is trained with Adam optimizer at $\eta = 10^{-3}$. This prevents random head weights from corrupting pretrained features.

Stage 2 - Full Fine-Tuning (15 epochs): All parameters are unfrozen and trained at $\eta = 10^{-5}$ with EarlyStopping (patience = 5) to prevent catastrophic forgetting of ImageNet representations.

Table III summarizes key hyperparameters across all three models.

TABLE III
TRAINING HYPERPARAMETERS COMPARISON

Parameter	ViT	Swin	IRNetV2
Optimizer	AdamW	Adam	Adam
LR (fine-tune)	3×10^{-5}	10^{-5}	10^{-5}
LR Scheduler	Cosine	None	None
Batch Size	32	32	32
Max Epochs	15+2	5+15	5+15
Label Smoothing	0.1	None	None
Grad Clipping	1.0	None	None
Weighted Sampler	Yes	No	No
Framework	PyTorch	PyTorch	TF/Keras

All models were trained on NVIDIA Tesla T4 GPU (15.6 GB VRAM) using CUDA 12.8 on the Kaggle Notebooks platform.

VI. EXPERIMENTAL RESULTS

A. Overall Classification Performance

Table IV presents the overall validation accuracy, macro-averaged F1-score, weighted F1-score, and macro-averaged AUC for all three architectures.

TABLE IV
OVERALL VALIDATION PERFORMANCE ON APTOS 2019 (733 IMAGES)

Model	Acc.	Macro F1	Wtd. F1	AUC
InceptionResNetV2	81.40%	0.7012	0.8121	0.921
Swin Transformer	83.20%	0.7198	0.8284	0.940
ViT-Base/16	**85.40%**	**0.7247**	**0.8482**	**0.951**

ViT-Base/16 achieves the highest validation accuracy of 85.40% and a macro F1-score of 0.7247, outperforming both the Swin Transformer and InceptionResNetV2. All architectures achieve AUC > 0.92, confirming strong discriminative capability.

B. Per-Class Classification Report (ViT-Base/16)

Table V presents the detailed per-class precision, recall, and F1-score for ViT-Base/16 on the 733-image validation set, obtained from the best model checkpoint.

TABLE V
PER-CLASS CLASSIFICATION REPORT – VIT-BASE/16

Class	Prec.	Recall	F1	Support
Mild NPDR	0.6800	0.4595	0.5484	74
Moderate NPDR	0.7386	0.8900	0.8073	200
No DR	0.9834	0.9834	0.9834	361
Proliferative DR	0.8125	0.6610	0.7290	59
Severe NPDR	0.6061	0.5128	0.5556	39
Accuracy		0.8540		733
Macro Avg	0.7641	0.7013	0.7247	733
Weighted Avg	0.8521	0.8540	0.8482	733

The model achieves near-perfect classification on No DR (F1 = 0.9834), which constitutes 49.3% of the dataset. Moderate NPDR achieves strong recall (0.89), benefiting from the largest minority class support (200 samples). Mild NPDR exhibits the lowest recall (0.4595) due to its visual similarity with both No DR and Moderate grades-a clinically recognized ambiguity at borderline lesion density. Severe NPDR has limited support (39 validation samples) which constrains reliable metric estimation.

C. Training Dynamics

Fig. **??** illustrates the epoch-by-epoch training dynamics of ViT-Base/16. The warmup phase (Epochs 1-3) drives rapid accuracy gains from 42.47% to 69.55% as the learning rate ramps from 3×10^{-6} to 3×10^{-5}. The cosine annealing phase (Epochs 4-15) achieves smooth convergence, reaching a peak validation accuracy of 85.40% at the extended Phase 2 checkpoint.

Table VI shows selected epochs from the ViT training log:

TABLE VI
VIT-BASE/16 TRAINING PROGRESS (SELECTED EPOCHS)

Epoch	LR	Tr.Loss	Tr.Acc	Val.Loss	Val.Acc
1	1.2×10^{-5}	1.4392	42.47%	1.0543	67.39%
3	3.0×10^{-5}	0.9673	69.55%	0.7710	79.67%
7	2.3×10^{-5}	0.7070	84.40%	0.7330	84.31%
12	5.3×10^{-6}	0.5015	94.64%	0.7698	84.58%
14	1.5×10^{-6}	0.4641	96.72%	0.7957	85.13%
15	1.0×10^{-6}	0.4494	97.51%	0.7931	84.45%
P2-E2	-	0.4437	97.51%	0.7927	**85.40%**

The persistent gap between training accuracy ($\sim$97.5%) and validation accuracy ($\sim$85.4%) at convergence suggests mild overfitting, managed by label smoothing and gradient clipping. For Swin Transformer, Stage 2 fine-tuning consistently improves over Stage 1, confirming that full backbone adaptation is necessary for the retinal domain shift from ImageNet.

InceptionResNetV2 exhibits stable convergence in both stages, with EarlyStopping preventing over-training.

D. Confusion Matrix Analysis

For ViT-Base/16, the confusion matrix reveals:

- **No DR** (Grade 0): Near-perfect recall (∼98%), as the absence of pathological features is unambiguous.
- **Mild NPDR** (Grade 1): Frequently confused with Moderate (Grade 2), reflecting the clinical overlap in microaneurysm density thresholds.
- **Severe NPDR** (Grade 3): Occasional misclassification as Proliferative DR due to similar hemorrhage patterns.
- **Proliferative DR** (Grade 4): Strong precision (0.8125) but moderate recall (0.661), suggesting some severe cases are under-identified.

The most clinically significant errors are Grade 1→Grade 0 misclassifications (missed mild DR) and Grade 3→Grade 2 misclassifications (under-staging of severe DR), both of which could delay clinical intervention.

E. ROC-AUC Analysis

One-vs-Rest (OvR) ROC curves show that ViT achieves AUC > 0.93 for all individual classes, with the highest AUC for No DR (≈ 0.99) and Moderate (≈ 0.96). The macro-averaged AUC of 0.951 confirms robust multi-class discrimination. Swin Transformer achieves a macro AUC of 0.940, slightly lower due to less aggressive augmentation and the absence of a learning rate scheduler. InceptionResNetV2 achieves a macro AUC of 0.921, reflecting the limitation of local receptive fields in capturing globally distributed DR lesion patterns.

F. GradCAM and GradCAM++ Explainability

Explainability maps were generated using GradCAM (InceptionResNetV2, Swin) and GradCAM++ (ViT) to identify diagnostically relevant retinal regions. For ViT, the target layer is `model.backbone.blocks[-1].norm1` (final transformer block's LayerNorm), with a ViT-specific reshape transform that removes the CLS token and reshapes the 196 patch token activations to a 14×14 spatial grid for overlay on the 224×224 input.

Table VII summarizes the clinically observed attention patterns per DR grade:

TABLE VII
GRADCAM ATTENTION PATTERNS VS. CLINICAL DR MARKERS

Grade	Clinical Marker	GradCAM Attention (ViT)
0 (No DR)	Absent lesions	Diffuse low-activation background
1 (Mild)	Microaneurysms	Focal perifoveal attention
2 (Mod.)	Exudates, hemorrhages	Attention on exudate clusters
3 (Sev.)	Venous beading, IRMA	Diffuse hemorrhage-region attention
4 (Prolif)	Neovascularization	Strong optic disc attention

For correctly classified samples, GradCAM++ maps consistently highlight pathology-bearing regions-perifoveal areas for mild DR, hard exudate clusters for moderate DR, and optic disc margins for proliferative DR. For misclassified samples (Mild confused as Moderate) the attention map shows diffuse activation without focused lesion localization, suggesting the model attends to non-diagnostic texture features. InceptionResNetV2's GradCAM (target: `conv_7b_ac` at 8×8 resolution) provides coarser spatial localization, while Swin Transformer's hierarchical GradCAM from the final stage offers intermediate spatial detail.

VII. COMPARATIVE ANALYSIS AND DISCUSSION

A. Accuracy vs. Parameter Efficiency

Fig. **??** summarizes the accuracy-vs-parameters trade-off. InceptionResNetV2 achieves 81.40% accuracy with only 54.3M parameters, making it the most parameter-efficient architecture. However, it sacrifices global contextual modeling due to CNN's local receptive field constraint. ViT-Base/16 achieves the highest accuracy (85.40%) with 86.2M parameters, while Swin Transformer achieves 83.20% with 87.0M parameters-slightly more parameters than ViT but with lower accuracy due to windowed (local) attention scope.

B. Local vs. Global Attention

The key architectural distinction driving performance differences is the scope of attention. CNNs (InceptionResNetV2) build receptive fields hierarchically through stacked convolutions, reaching global scope only in the deepest layers. ViT achieves global self-attention from the very first transformer block, enabling direct modeling of long-range spatial dependencies between lesions in opposite retinal quadrants. This is particularly beneficial for Severe and Proliferative DR where pathology is spatially distributed. Swin Transformer achieves a middle ground through shifted windows, providing cross-window context modeling at linear computational cost.

C. Training Strategy Comparison

The single-phase AdamW fine-tuning with cosine annealing (ViT) outperforms the two-stage Adam training (Swin, InceptionResNetV2). Several factors contribute:

1) AdamW's decoupled weight decay is better suited for transformers than standard Adam.
2) Cosine annealing provides smooth LR decay, preventing sharp loss oscillations in the late training phase.
3) Label smoothing (ViT only) acts as a soft regularizer that reduces overconfidence on ambiguous grade boundaries-clinically important given inter-grader variability.
4) WeightedRandomSampler (ViT only) ensures balanced gradient contributions from all five classes per batch.

D. Augmentation Impact

ViT uses the richest augmentation pipeline (7 transforms), which is critical for a 3,662-image dataset. Swin uses minimal augmentation (flip + rotation), which may explain its lower

performance despite architectural advantages. InceptionResNetV2 applies zoom augmentation (Keras-native) that simulates camera distance variation-a practical consideration for fundus imaging.

E. Class-Level Challenges

All three models share consistent class-level difficulties:

- **Mild NPDR**: Lowest recall across all models due to sparse microaneurysm patterns visually overlapping with No DR.
- **Severe NPDR**: Low validation support (39 samples) limits reliable metric estimation and model training signal.
- **No DR**: Highest performance across all models due to dominant class representation and absence of pathological features.

F. Clinical Deployment Considerations

- **Accuracy**: ViT $>$ Swin $>$ InceptionResNetV2
- **Inference Speed**: InceptionResNetV2 $>$ Swin $\approx$ ViT
- **VRAM Requirement**: ViT ($\sim$4 GB at BS=1) $>$ Swin $>$ InceptionResNetV2
- **Explainability**: GradCAM++ (ViT) provides the finest spatial resolution for patch-level localization; GradCAM (InceptionResNetV2) is coarser (8×8 spatial resolution).
- **Deployment**: InceptionResNetV2 (TF/Keras .h5) is most portable; ViT and Swin (.pth) require PyTorch with `timm`.

Table VIII provides a comprehensive head-to-head comparison.

TABLE VIII
COMPREHENSIVE ARCHITECTURE COMPARISON

Property	IRNetV2	Swin	ViT
Attention Type	Local (CNN)	Local+Global	Global
Parameters	54.3M	87.0M	86.2M
Validation Acc.	81.40%	83.20%	**85.40%**
Macro F1	0.7012	0.7198	**0.7247**
Macro AUC	0.921	0.940	**0.951**
Training Strategy	2-stage	2-stage	1-stage
LR Scheduler	None	None	Cosine
Explainability	GradCAM	GradCAM	GradCAM++
Framework	TF/Keras	PyTorch	PyTorch
Complexity	$O(n^2)$ CNN	$O(n)$	$O(n^2)$ ViT

VIII. System Architecture and Deployment

The trained models are integrated into a Flask-based web application for clinical screening demonstrations. The backend (`app.py`) loads the best model checkpoint via `model_loader.py`, performs preprocessing through `utils/preprocessing.py`, and returns the predicted DR grade with GradCAM visualization generated by `gradcam.py`. The frontend provides an image upload interface (`upload.html`) with result display (`result.html`) showing the predicted grade and heatmap overlay.

The deployment pipeline supports:

- Single-image inference with confidence scores for all 5 DR grades
- GradCAM heatmap generation for clinical explainability
- Multi-model switching via `configs/deployment_config.yaml`
- GPU-accelerated inference with CUDA auto-detection

IX. Limitations and Future Work

Dataset Size: The 3,662-image APTOS 2019 dataset is relatively small for transformer architectures that typically require large-scale pretraining data. Training on larger datasets (EyePACS: 88,702 images; Messidor-2: 1,748 images) would likely improve generalization.

Resolution Constraint: All models operate at 224×224 pixels. Clinical fundus images are often 2×2 megapixels, and downsampling may lose sub-millimeter microaneurysm detail critical for Grade 1 detection.

No Independent Test Set: Evaluation uses a validation split from the same data distribution. An independent multi-center test set is necessary for unbiased clinical validation.

Class Imbalance: Despite WeightedRandomSampler (ViT), all models exhibit degraded performance on Severe NPDR (Grade 3) with limited support. Synthetic augmentation via SMOTE, CycleGAN, or diffusion models could generate clinically plausible minority class samples.

Future Directions:

- Ensemble of ViT + Swin with majority voting or learned fusion weights
- Knowledge distillation from ViT to a compact CNN for edge deployment
- Contrastive pretraining on unlabeled fundus images for better domain adaptation
- Multi-task learning for simultaneous DR grading and lesion segmentation
- Temporal modeling for longitudinal progression prediction

X. Conclusion

This paper presented a comprehensive comparative study of three deep learning architectures-ViT-Base/16, Swin Transformer Base, and InceptionResNetV2-for five-class diabetic retinopathy grading on the APTOS 2019 dataset. ViT-Base/16, trained with end-to-end AdamW optimization, cosine annealing scheduling, label smoothing, and weighted sampling, achieves the highest validation accuracy of **85.40%** with a macro F1-score of 0.7247 and macro AUC of 0.951. Swin Transformer achieves 83.20% accuracy, and InceptionResNetV2 achieves 81.40%.

The global self-attention mechanism of ViT is particularly well-suited for DR grading where pathological lesions are spatially distributed across the retinal field. GradCAM++ visualizations confirm that ViT's attention maps align with clinically validated lesion markers-perifoveal microaneurysms for Mild DR, hard exudate clusters for Moderate DR, and optic disc neovascularization for Proliferative DR-providing

clinical interpretability alongside competitive classification performance.

These results demonstrate that transformer-based architectures even with modest dataset sizes ($\sim$3,700 images) can achieve competitive DR grading performance through careful hyperparameter optimization, rich augmentation, and pretraining on large-scale natural image datasets. The integration of GradCAM explainability into the deployment pipeline supports clinical adoption by providing transparent, visually verifiable predictions for ophthalmologist review.

ACKNOWLEDGMENT

The authors thank the APTOS 2019 Asia Pacific Tele-Ophthalmology Society for making the fundus image dataset publicly available. Computational resources were provided through Kaggle Notebooks with Tesla T4 GPU access.

REFERENCES

[1] International Diabetes Federation, "IDF Diabetes Atlas, 10th edition," 2021. [Online]. Available: https://www.diabetesatlas.org

[2] World Health Organization, "Blindness and vision impairment," WHO Fact Sheet, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment

[3] 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.

[4] D. S. W. Ting et al., "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *JAMA*, vol. 318, no. 22, pp. 2211-2223, 2017.

[5] M. D. Abràmoff et al., "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," *Investigative Ophthalmology & Visual Science*, vol. 57, no. 13, pp. 5200-5206, 2016.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.

[7] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.

[8] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF ICCV*, pp. 10012-10022, 2021.

[9] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI*, 2017.

[10] V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402-2410, 2016.

[11] M. D. Abràmoff et al., "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices," *NPJ Digital Medicine*, vol. 1, no. 1, pp. 1-8, 2018.

[12] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962-969, 2017.

[13] B. Graham, "Kaggle diabetic retinopathy detection competition report," University of Warwick, 2015.

[14] N. Sikder, M. S. Masud, A. K. M. B. Hossain, and M. A. Bhuiyan, "Severity classification of diabetic retinopathy using an ensemble learning algorithm through analyzing retinal images," *Symmetry*, vol. 13, no. 4, p. 670, 2021.

[15] S. Qummar et al., "A deep learning ensemble approach for diabetic retinopathy detection," *IEEE Access*, vol. 7, pp. 150530-150539, 2019.

[16] X. Wang et al., "Self-attention based CNNs for diabetic retinopathy grading using fundus image," in *Proc. IEEE ISBI*, 2021.

[17] X. Sun, J. Xu, and J. Ma, "Vision transformer for diabetic retinopathy grading," in *Proc. Int. Conf. on Medical Image Analysis and Computer-Aided Diagnosis*, 2021.

[18] B. Gheflati and H. Rivaz, "Vision transformers for classification of diabetic retinopathy," in *Proc. IEEE EMBC*, pp. 1988-1991, 2022.

[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE/CVF ICCV*, pp. 618-626, 2017.

[20] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE WACV*, pp. 839-847, 2018.

[21] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. IEEE/CVF CVPR*, pp. 782-791, 2021.

[22] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[23] R. Wightman, "PyTorch Image Models (timm)," GitHub, 2019. [Online]. Available: https://github.com/rwightman/pytorch-image-models

[24] Asia Pacific Tele-Ophthalmology Society, "APTOS 2019 Blindness Detection," Kaggle Competition, 2019. [Online]. Available: https://www.kaggle.com/c/aptos2019-blindness-detection