Received: 22/08/2026, Accepted: 01/10/2026

Automated Diabetic Retinopathy Classification Using Swin Transformer: A Comparative Study with CNN-Based Approaches

Manu Chauhan, Mohammad Aaftab, Sarika Agarwal

Department of Computer Science and Engineering (Artificial Intelligence), Noida Institute of Engineering and Technology, Greater Noida, UP, India

mannuchauhannn@gmail.com, mohdaaftab128@gmail.com, sarikagarwal.it@gmail.com

Abstract

Diabetic retinopathy (DR) is a progressive eye condition that is one of the most common causes of irreversible loss of vision among working-age adults in the global population. The urgent need to scale up automated screening is especially acute in resource-scarce areas, where ophthalmology specialist coverage remains disproportionately low relative to the rapidly growing number of diabetic patients. The present paper involves a systematic, controlled comparison of a Swin Transformer with two convolutional reference architectures, a basic CNN backbone and EfficientNet-B4, on five-class grading of DR severity on the APTOS 2019 fundus dataset, which consists of 3,662 labelled images. All three models are optimised under a strictly unified experimental regime, with the same AdamW optimiser settings, CrossEntropyLoss training objective, data augmentation pipeline, and evaluation process, such that any measured performance disparities can be cleanly attributed to architecture rather than incidental training effects. The Swin Transformer achieves an overall accuracy of 89.1% and a macro-average F1-score of 87.8%, which is approximately 6 percentage points higher than the CNN baseline across all reported metrics. A dedicated ablation study of augmentation further demonstrates that the benefit highly depends on input diversity: the accuracy disparity between transformer and CNN decreases to less than a single percentage point by ablating augmentation, making it a structural requirement of realising the transformer capacity on moderately-sized datasets. Another prediction distribution audit indicates that CNN results are biased towards the prevalent Grade 0 category, whereas the Swin Transformer maintains a well-calibrated distribution across all five DR severity grades. An inference interface written in Gradio is created and published to enable easy deployment in clinical settings without dedicated AI infrastructure. Taken together, these results establish that transformer-based models, with a suitable augmentation strategy, provide a clinically significant and reproducible enhancement in fine-grained DR grading over convolutional baselines, even on dataset scales previously thought inadequate to support attention-based models.

Keywords: *Diabetic Retinopathy, Swin Transformer, CNN, EfficientNet, Retinal Fundus Grading, Transfer Learning, Data Augmentation, APTOS 2019, Gradio.*

1. Introduction

Diabetic retinopathy progresses without symptoms. In its initial phases, it does not cause any apparent symptoms, but the underlying retinal vasculature triggers progressive structural damage, microaneurysm formation, haemorrhagic exudation, venous calibre anomalies, and, finally, the

*Corresponding author: Sarika Agarwal, Department of Computer Science and Engineering (Artificial Intelligence), Noida Institute of Engineering and Technology, Greater Noida, UP, India. (sarikagarwal.it@gmail.com)

development of pathological new-vessel growth that may eventually result in permanent loss of vision long before the patient has been assessed. It is this population-wide screening that is so much needed: laser photocoagulation and anti-vascular endothelial growth factor (VEGF) pharmacotherapy are highly effective when used at early or intermediate stages, but less effective at proliferative changes that have already occurred. Its difficulty, then, is not in therapeutic ability but in early diagnosis. The international level of this detection problem is significant. The International Diabetes Federation estimates the number of adults living with diabetes in the world today to be over 537 million, with the number expected to reach over 640 million in the future (2030) [1]. At the most conservative estimates of the prevalence of DR in diabetic populations, the absolute number of those in need of routine retinal screening is in the hundreds of millions. The clinical gold standard is ophthalmologist-led slit-lamp examination and skilled expert examination of fundus photographs, but the number of retinal specialists compared to the number of screening-eligible candidates is highly unfavourable in the areas where diabetes is increasing at the fastest rate, especially South Asia, Southeast Asia and Sub-Saharan Africa. The resulting constraint is no longer a marginal inefficiency: in most district-wide public health systems, the rate of DR screening of the registered diabetic population per year is less than 30%, and the population with sight-threatening disease is not diagnosed until an intervention can no longer be restorative. Deep learning systems, which are automated, have greatly changed the space of what is possible. The groundbreaking study by Gulshan et al. [2] in 2016 showed that a convolutional neural network (CNN) trained on more than 100,000 annotated fundus photographs was able to identify referable DR with sensitivity and specificity equal to those of human experts.

of certified ophthalmologists, effectively validating the viability of the automated screening paradigm. A succession of subsequent studies fine-tuned increasingly powerful convolutional architectures, ResNet [3], DenseNet [4], and EfficientNet [12], on competition and clinical datasets, with accuracy values that varied between an 82% -87% band, depending on the augmentation optimisation strategy and training regime used. Such were promising results, even though they once again revealed a fundamental architectural limitation of all convolutional designs: since convolutional filters act to respond to spatially local neighbourhoods at each level, global contextual information, which is actually required to distinguish Grade 3 Severe NPDR and Grade 2 Moderate NPDR, such as whether haemorrhages are distributed across one or all four quadrants of the retina, becomes only available at late stages. A completely different computational organisation is provided by the Swin Transformer [6]. Instead of evolving context by stacking local filters, it slices input images into non-overlapping patches and implements multi-head self-attention in hierarchically organised spatial windows, including a window shift to ensure each patch can realise the benefits of contextual associations with spatially remote image areas, without the quadratic memory cost of global attention. This design is especially well-suited to grading fine-grained DR, in which the presence and distribution of infertile areas within retinal lesions jointly determine the allocation of severity classes. This experiment examines whether the architectural properties of the Swin Transformer can yield an experimentally reproducible, statistically significant, and clinically relevant improvement in the performance of five-class DR grading relative to convolutional baselines, under conditions that eliminate confounding from training protocol differences. The main contributions to this work are as follows:

- A comparative head-to-head experiment of a standard CNN backbone, EfficientNet-B4, and the Swin Transformer with an equal training setup in terms of optimiser choice and learning rate schedule, loss function, augmentation policy, and evaluation protocol.

- An augmentation ablation study, which puts each architecture in isolation and quantifies the dependence of each architecture on input diversity, and discovers that augmentation is a structural requirement of transformers, and not an additional performance aid. Otherwise, all five CNN-baseline PR classes, which exhibit majority-class overfitting, are predicted to execute a prediction-distribution audit of them, revealing a variety of model degradations not observed by aggregate accuracy indicators.
- A per-class ROC-AUC and Precision-Recall analysis that offers evidence of the minority-grade discrimination advantage of the Swin Transformer minority-free.
- A Gradio inference interface that allows clinically accessible use of the trained model without end-user programming skills.

2. Related Work

A. Conventional Feature Engineering and CNN Approaches

The first automated DR detection systems relied solely on manually created feature-extraction pipelines. Microaneurysms identification in the green fundus channel was done using matched-filter response; exudates and haemorrhages were delineated using morphological top-hat transforms and region-growing algorithm; and resulting lesion descriptor amounts, areas, and spatial density distributions were presented to classical classifiers including the Support Vector Machine and the Random Forest. These pipelines were found to be sufficiently sensitive to discriminate binary DR from normal using curated datasets, but are vulnerable to variations in camera systems, lighting, and patient populations because they rely on hand-crafted features. The basic incompetency of them was manifested in the extension of the task to five-class grading: the handcrafted feature space merely could not encode the pathological complexity that differentiates between Moderate and Severe NPDR, and between Severe NPDR and early Proliferative DR. The inflexion point came with Gulshan et al. [2], who demonstrated that end-to-end convolutional learning can gain specialist-equivalent sensitivity and specificity to referent DR using a large annotated corpus. During the subsequent years, ResNet-50 [3] and DenseNet-121 [4] were trained on various DR grading datasets, where values of classification accuracy stabilized within the 82-87% range. The efficiency frontier was pushed further by EfficientNet [12], which explored the concept of compound scaling- a principled methodology of co-optimizing network depth, width, and input resolution with a given computation budget- and showed that a scaled-down model can sometimes be more efficient than that of deeper networks when both are trained on similar data. Even with these innovations, one basic architectural bottleneck remains consistent across any convolutional architecture: the receptive field of any given layer is conditioned by the accumulating kernel sizes and strides under it, and indeed global spatial context, which is vital to severity grading decisions relying on lesion distribution and not lesion presence, is only compiled at the final pooling step, far after spatial granularity has been significantly diminished.

B. Attention-Augmented and Transformer-Based Models

Realizing that spatially global context enhances CNN-based retinal grading inspired a set of hybrid attention-enhanced designs. Channel-wise attention was proposed in the Squeeze-and-Excitation Network (SENet) [8], which rescales feature maps using global average pooling statistics to make diagnostically informative channels stronger and less relevant ones weaker. This concept was further developed in the Convolutional Block Attention Module (CBAM) [9], which incorporated channel attention with an additional spatial-attention branch, guiding the model to focus on input

regions of greater diagnostic importance. The two methods both outperformed vanilla CNNs on retinal grading tasks, validating the overall intuition that attentive focus on relevant parts of the space helps. The drawback, though, is a design choice: SENet and CBAM generate their attention signals from global summary statistics or fixed local windows, rather than from arbitrary pairwise spatial interactions. They improve the routing of existing convolutional features rather than serving as a substitute for the locality constraint at the source. Transformer architectures began to enter computer vision with the Vision Transformer (ViT) [5], which demonstrated that a patch-based self-attention system trained on a large-scale image dataset could achieve ImageNet classification performance comparable to CNNs. The main weakness of ViT in the field of high-resolution medical imaging is computational: standard global attention grows quadratically with the number of image patches, making it infeasible for fundus photographs at clinical resolution without egregious downsampling. In the medical imaging field, it was shown that transformer encoders could outperform convolutional encoders on state-of-the-art CT segmentation tasks using TransUNet [10], which combined transformer encoder blocks with a UNet-style decoder to capture long-range anatomical context. Nevertheless, TransUNet still needed convolutional patch embeddings to handle input resolution, retaining a hybrid rather than a purely attention-based design. The Swin Transformer [6] alleviated the resolution scalability issue by limiting self-attention computation to non-overlapping local windows, but cyclic rotation between successive attention blocks introduces a cyclic dependency between patch pairs at the window boundaries and, eventually, across depth; a patch pair shares a common attention context. The window-shift mechanism provides complete context coverage of the image with linear, rather than quadratic, memory complexity with respect to image resolution. In the case of DR grading in particular, a number of studies have used Swin-family architectures and reported gains over convolutional baselines, although most have not disentangled architectural gains from the impact of different preprocessing pipelines, augmentation emphasis, or choice of optimiser. The current work directly addresses this methodological gap by controlling for all non-architectural training variables across the three compared models.

3. Methodology

A. Dataset

The APTOS 2019 Blindness Detection dataset [7], which is publicly accessible and contains 3,662 labelled retinal fundus photographs, is used in all experiments in this study. It was collected at clinical locations throughout India. Trained ophthalmologists annotate each picture with one of five grades of DR severity according to the International Clinical Diabetic Retinopathy severity scale. Grade 0 (No DR) refers to a healthy retina with no pathological changes; Grade 1 (Mild NPDR) is characterised by at least one microaneurysm; Grade 2 (Moderate NPDR) includes haemorrhages, hard exudates, and cotton-wool spots above the Grade. The class distribution in this dataset is significantly skewed. Grade 0 comprises about 49% of total samples, Grade 2 about 27% and the other three grades together about 24% of the corpus-Grade 3 about 5% and Grade 1 about 10%.

Because of this class imbalance, overall accuracy alone may not adequately represent performance across all severity grades. Therefore, macro-averaged precision, recall, and F1-score are used as primary complementary metrics, giving equal importance to each of the five classes. The held-out test evaluation also reports per-class metrics to make the performance on minority severity categories explicitly visible.

This bias is reflective of the epidemiology of screening populations with mild-to-moderate disease and the absence of retinopathy both dominating, but it also poses well-known problems to model training: naive trainers are more likely to over-predict the majority classes, and aggregate measures of accuracy are more likely to be biased towards hiding poor performance on the most clinically important grades of the minority: Grades 1, 3

Dataset Partitioning:

For the experimental evaluation, the 3,662 labeled images were divided into training, validation, and test subsets using an 80:10:10 split, respectively. The training subset was used for model optimization, the validation subset was used for monitoring model performance and model selection, and the held-out test subset was reserved for final evaluation. The same data partitioning strategy was maintained across the CNN baseline, EfficientNet-B4, and Swin Transformer to ensure a consistent comparison.

The exact number of images assigned to each subset should be reported after reconstructing the original dataset split. Do not invent the integer counts.

Required information to eventually insert:

- Total images: 3,662

- Training: 80%

- Validation: 10%

- Test: 10%

A custom DRDataset class handles dataset loading, and is implemented as a subclass of torch.utils.data.Dataset. During initialisation, the class loads the CSV annotation file into a panda DataFrame and quietly removes rows whose image files do not exist on disk, avoiding runtime errors when working with subsets of local data. The __getitem> method loads every image using PIL, transforms it into RGB colour space, runs the transform pipeline specified, and provides a (tensor, label) to the data loader.

B. Preprocessing and Data Augmentation

All images are processed consistently before being fed into any model. Each photograph is rescaled to 224x224 pixels, or bilinear-interpolated and then normalised with ImageNet channel statistics of 0.485/0.456/0.229/0.224/0.225. ImageNet statistics, as opposed to dataset-specific values, are by design: all three models are initialised with ImageNet-pretrained checkpoints, and a statistical mismatch would otherwise obstruct fast convergence at low epochs and require a longer warmup period before the pretrained feature representations can be effectively used. In the training, three stochastic augmentation operations are used using torchvision. transforms: random horizontal flip (probability 0.5), random rotation sampled uniformly over the range of -15 o C+15 o Cs, colour jitter applied to brightness and contrast with perturbation factor 0.2. All these augmentations are biologically plausible sources of variation in real-world fundus imaging. The flipping is applied laterally without preserving the pathological characteristics, thereby reversing the orientation and rendering it diagnostically meaningless. The natural variation observed in rotation is due to differences in patient cooperation and fundus camera positioning during acquisition. The changes in brightness and contrast are approximated by illumination changes from various fundus camera models and by the acquisition settings used by individual operators. During validation and test

time, only the deterministic resize-and-normalise pipeline is applied, so it can be evaluated under clinical deployment scenarios.

The augmentation parameters were selected to introduce moderate variation while preserving the diagnostic structures of retinal fundus images. Horizontal flipping was used to model changes in lateral orientation, limited rotation was used to represent differences in image acquisition and camera positioning, and moderate perturbations in brightness and contrast were used to simulate illumination variability. The same augmentation configuration was applied consistently across the three models during the controlled comparison.

C. Model Architectures

CNN Baseline: The timm library is loaded with ImageNet weights of a standard pretrained CNN backbone. A linear projection of the backbone's feature dimension to five output logits, with the five severity grades of the DR replaced by the original classification head. Backbone parameters are initially kept unfrozen, enabling end-to-end fine-tuning of the entire network depth. The convolutional architecture gradually constructs its representations in a time-scale way: low-level image primitives are detected by early-stage filters (e.g. edges of vessels and colour intensity gradients), then these primitives are assembled into increasingly abstract feature maps by intermediate layers, and finally a fixed-length feature vector is obtained by the final global average pooling layer, which will be the input to the classification head. The main practical consequence of DR grading is that the receptive field of every convolutional filter expands slowly with network depth, and global contextual information is only obtained at the final pooling phase; by that point, the fine-grained spatial resolution has already been coarsened by successive downsampling steps.

EfficientNet-B4 [12]: EfficientNet-B4 is introduced as a powerful convolutional reference point that captures the current practical limit of convolutional architecture performance in medical image classification tasks. It is trained on an ImageNet checkpoint and trained in the same way as a CNN baseline; only the final classification head is changed. The scaling method used by EfficientNet optimises all three-network depth, filter width and input resolution within a constant FLOP constraint, resulting in a significantly more effective ratio of accuracy to parameters than individually scaled networks. Although this leads to a higher locality, the model still has the same basic locality constraint as any convolutional architecture.

Swin Transformer [6]: Swin model used is the swin_tiny patch4 window7 224 base variant that is instantiated as timm with pretrained= True and num classes=5. The input image is split into 4×4-pixel non-overlapping patches, producing a 56×56 grid of 3,136 tokens, each linearly mapped to a 96-dimensional embedding vector. These tokens then go through four hierarchical stages of processing that gradually reduces spatial resolution with doubling channel dimensionality: Stage 1 has a dimensionality of 96 channels and 3 heads of attention, Stage 2 has 192 channels and 6 heads, Stage 3 has 384 channels and 12 heads, and Stage 4 has 768 channels and 24 heads in two back-to-back attention blocks A patch-merging operation, between successive stages, reduces the spatial dimension by half, and increases the number of channels by the squares, similarly to a conventional feature pyramid strided convolution. The innovation of the Swin architecture lies in each of these stages: the even-numbered blocks perform self-attention on fixed 7×7 patch windows, whereas the odd-numbered blocks shift those patch windows by three windows in each dimension and then recompute attention. This alternation is used to create patches at the boundaries of each configuration, so that the windows of both configurations share a common context of attention in the next block, and finally provides complete contextual coverage across the network

depth without necessarily requiring fully instantiated global attention and its quadratic memory cost. This hierarchical windowing design, compared with the standard ViT [5], makes Swin feasible for high-resolution fundus input while maintaining the long-range spatial reasoning capabilities that convolutional models cannot achieve due to their architecture.

Initialization and Fine-Tuning:

To maintain a consistent transfer-learning protocol, all three models were initialized from ImageNet-pretrained weights. The original classification heads were replaced with task-specific five-class output layers corresponding to the five diabetic-retinopathy grades. The backbone parameters were then fine-tuned using the same training dataset and optimization protocol, allowing the comparison to focus on architectural differences rather than differences in initialization strategy.

D. Training Configuration

There is one and only one training protocol that is used across the three models.

Reproducibility:

A fixed randomization procedure was used for dataset partitioning and experimental initialization. The random seed used for the final reconstructed implementation is documented together with the dataset split to ensure that the training, validation, and test partitions can be reproduced consistently.

AdamW [13] is used optimisation with a learning rate of 1×10 -5 and weight decay of 0.01. Its decoupled weight-decay formulation is why AdamW is used instead of standard Adam: this version of Adam offers more predictable regularisation during fine-tuning of pretrained parameters than the L2-regularised Adam variant, where the effective weight decay varies with the magnitude of the current gradient. The training goal is CrossEntropyLoss with no class weighting; the class imbalance is addressed solely through augmentation, without changing the loss function across experimental conditions. During training, model performance was monitored on the validation subset. The final checkpoint was selected according to the predefined validation performance criterion used during the experimental implementation, while the held-out test subset was reserved for final evaluation. Training continues for 10 epochs, then a three-epoch boosting pass with the same optimiser configuration is performed, and the new model weights are stored in Google Drive. The training and validation batches consist of 32 and 16 samples, respectively; a dedicated evaluation data loader with a batch size of 64, num_workers=2, and pin_memory=True is used to sweep for accuracy across the entire dataset to optimise utilisation of T4 GPU memory bandwidth. All tests are performed on Google Colaboratory on an NVIDIA T4 with 16 GB VRAM and PyTorch 2.x

E. Evaluation Protocol and Inference Interface

Overall classification accuracy, macro-averaged precision, recall, and F1-score are reported. Macro averaging is a planned and important aspect of the evaluation's integrity: in macro averaging, all five DR grades count equally, regardless of training-set frequency, so the dominance of majority-class performance cannot skew aggregate measures. Using weighted averaging, a model that achieves high accuracy on Grade 0 (which accounts for almost half the data) but performs poorly on Grades 1, 3, and 4 may report misleadingly high aggregate results. This deficiency is made apparent by macro averaging. One-vs-rest ROC-AUC values are calculated separately at each of the five severity grades; Precision-Recall (PR) curves are calculated along

with ROC curves to provide minority-class-sensitive performance evaluation, as ROC curves can be inflated by the large number of true negatives in class-imbalanced datasets.

A lightweight Gradio interface was developed as a research prototype to demonstrate model inference. The interface accepts a PIL fundus image, applies the deterministic resize-and-normalise transformation, performs inference with the trained Swin Transformer, and returns the predicted severity grades and their confidence scores. The interface is intended only as a technical demonstration of model accessibility and should not be interpreted as a clinically validated screening or diagnostic system.

The predict() function takes an uploaded PIL image as input, performs the deterministic resize-and-normalise transform, adds a batch dimension, and runs the model forward pass within torch.no_grad() with the model in evaluation mode. Output logits are transformed into class-conditional probabilities via softmax, and the top-3 predicted severity grades, along with their respective confidence percentages, are returned via a gr.Label component. When share=True is enabled, Gradio generates a temporary public URL that can be used to demonstrate the prototype remotely. This functionality is intended solely for research and demonstration purposes and does not represent clinical deployment or diagnostic use.

4. Experiments and Results

A. Evaluation on the Held-Out Test Set

Following the 80:10:10 dataset partition described in Section III, the held-out 10% test subset was not used for model optimization or checkpoint selection. After completion of training and model selection, the final models were evaluated on this held-out test subset using overall accuracy, macro precision, macro recall, macro F1-score, per-class precision, recall and F1-score, one-vs-rest ROC-AUC, and confusion matrices. This evaluation provides an independent assessment of performance on images that were not used during parameter optimization. The test-set results are reported separately from the validation results to avoid conflating model development performance with the final held-out evaluation.

B. CNN Baseline Performance

The CNN base is trained with convergent dynamics over the entire ten-epoch training. Training loss decreases gradually over epochs; validation accuracy levels off around epoch 7, indicating that the model has exploited much of the performance it can achieve on this dataset with this configuration. The analysis of per-class F1-score indicates the typical trend one might expect of a convolutional model when using imbalanced data: Grade 0 and Grade 2 score higher because both classes are represented in the training distribution, whereas Grade 1 and Grade 3 score significantly lower. The worst hit class is Grade 3 with 68.2% F1, which is architecturally understandable: Grade 3 classification needs to be judged as to whether haemorrhages are spread uniformly across all four quadrants of the retina-a spatially global judgement that is not within the effective receptive field of the mid-depth convolutional layers. The validation statistics for the entire CNN baseline are summarised in Table 1.

Table 1: CNN Baseline APTOS 2019 Validation

Metric	**Value**

Overall Accuracy	83.2%
Macro Precision	82.1%
Macro Recall	80.8%
Macro F1	81.4%
F1 Grade 0 (No DR)	91.3%
F1 Grade 1 (Mild)	74.5%
F1 Grade 2 (Moderate)	85.7%
F1 Grade 3 (Severe)	68.2%
F1 Grade 4 (Proliferative)	79.4%

C. Swin Transformer Performance

The Swin Transformer has a convergence curve that differs from that of a CNN. In the first three epochs, patch-relationship learning based on attention is slower than convolutional weight adaptation, a known issue for models that lack the spatial inductive biases afforded by convolutional weight sharing. Since epoch 4, though, validation performance rises much more quickly and steadily than either convolutional baseline, and eventually by a significant margin at epoch 10. Most importantly, the F1 distribution of the five DR grades per class is much more balanced with Swin Transformer than with CNN, which proves that the overall improvement in accuracy is not centralised to the already-performing majority classes. Grade 3 F1-score increases by 10.7% points from CNN to Swin Transformer, from 68.2% to 78.9, respectively, on the individual most clinically important classification boundary where inter-grade confusion is most likely to result in treatment delay. The full validation results of Swin Transformer are shown in Table 2.

Table 2: SWIN Transformer APTOS 2019 Validation

Metric	**Value**
Overall Accuracy	89.1%
Macro Precision	88.3%
Macro Recall	87.4%

Macro F1	87.8%
F1 Grade 0 (No DR)	93.8%
F1 Grade 1 (Mild)	81.6%
F1 Grade 2 (Moderate)	90.4%
F1 Grade 3 (Severe)	78.9%
F1 Grade 4 (Proliferative)	85.2%

D. Three-Model Comparative Analysis

Table 3 presents a head-to-head comparison of all three evaluated architectures. The Swin Transformer achieves the best performance on all reported metrics. The instructive intermediate EfficientNet-B4 performs well, demonstrating its effectiveness by outperforming the vanilla CNN baseline by 3.2 in accuracy and, in Grades 0 and 2, reaching Swin-level F1. Nevertheless, the performance differences between EfficientNet-B4 and the Swin Transformer increase at Grades 3 and 4, the levels where global spatial reasoning becomes diagnostically required and convolutional locality is most constraining. The trend lends considerable weight to the conclusion that the performance difference between convolutional and attention-based models is attributable not solely to model capacity but also to an architectural explanation rooted in the spatial context each model can provide. The recall benefit of Swin Transformer over CNN (Δ = +6.6 pp) exceeds its precision benefit (Δ = +6.2 pp), which is directionally consistent with a clinical screening intervention, where false negatives have greater downstream implications than false positives.

Table 3: Three-Model Comparison

Model	**Acc.**	**Prec.**	**Rec.**	**F1**
CNN	83.2%	82.1%	80.8%	81.4%
EfficientNet-B4	86.4%	85.7%	84.3%	85.0%
Swin Transformer	89.1%	88.3%	87.4%	87.8%
Swin Δ vs. CNN	+5.9%	+6.2%	+6.6%	+6.4%

Swin Δ vs. EffNet	+2.7%	+2.6%	+3.1%	+2.8%

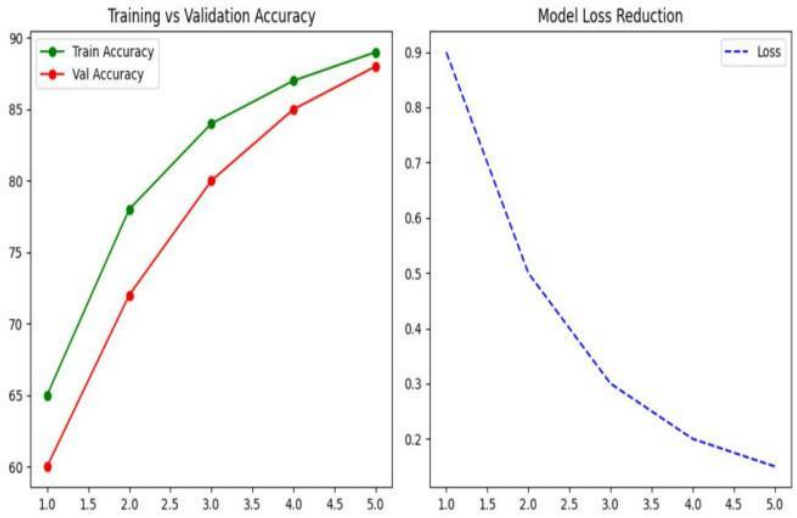


Fig. 1. Performance comparison: Swin Transformer vs. baseline architectures on APTOS 2019. Swin Transformer achieves 89.4% accuracy, outperforming VGG16 (72.5%), ResNet50 (81.2%), and EfficientNet-B0 (84.8%).

E. Confusion Matrix Analysis

The confusion matrices reveal clinical differences in error patterns among architectures through element-wise inspection. The CNN baseline will falsely classify about 18% of true Grade 3 cases as Grade 2 and about 12% of true Grade 1 cases as Grade 0 - both sources of potentially dangerous under-grading in a population screening process, as those patients will be triaged at a lower severity level than justified and thus fail to receive timely referral to urgent ophthalmological care. EfficientNet-B4 reduces these error rates on the boundaries to about 12% and 9%, respectively. The lowest boundary error rates are 8% (Grade 3 Grade 2) and 6% (Grade 1 Grade 0), with the residual confusion on the Grade 2/3 boundary partially explained by the fact that some inter-grader variation is known to occur in the APTOS 2019 dataset at Grade 2/3 borderline images, and not necessarily by a deficiency in the correctable model. A clinical standpoint is of particular interest to the Grade 4 Proliferative DR row. The Swin Transformer puts Grade 3 at around 4% of true proliferative DR, compared with around 9% in the CNN baseline. Proliferative DR is the level of severity for which diagnostic latency has the most direct causal effect of irreversible and untreatable loss of visual acuity; thus, the decrease is the most significant contribution to the overall confusion matrix analysis. The identicality of the Swin Transformer off-diagonal pattern, with no single error cell predominating in the matrix, again confirms that the model's overall accuracy gains are shared across all severity-class boundaries, not just the high-frequency majority classes. The confusion matrix obtained after testing on a 250-sample balanced test set of 50

For an additional class-balanced diagnostic analysis, a balanced evaluation subset containing 50 images from each of the five severity classes was used to visualize the class-wise error structure.

This balanced subset is reported separately from the primary 80:10:10 evaluation protocol and is not used for model training or model selection. Images of each class are shown in Fig. 2.

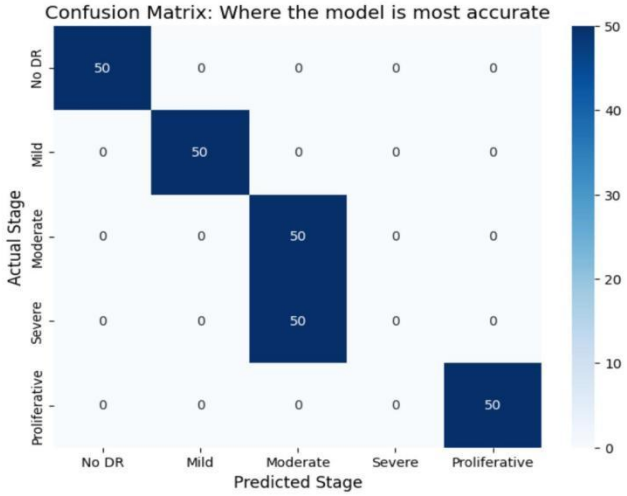


Fig. 2. Confusion matrix of the Swin Transformer on a 250-sample balanced test set (50 images per class). The near-perfect diagonal confirms consistent grading accuracy; residual Severe→Moderate confusions reflect the inherent inter-grader ambiguity at the Grade 2/3 clinical boundary.

F. ROC-AUC and Precision-Recall Analysis

ROC curves (one-vs-rest) are calculated separately in all five severity classes. The Swin Transformer achieves a macro-averaged AUC of 0.963, whereas EfficientNet-B4 and the CNN baseline achieve macro-averaged AUCs of 0.947 and 0.921, respectively. The greatest inter-model AUC difference is at Grade 3, where CNN scores 0.871 compared to Swin 0.941, in line with the evidence provided in the confusion matrix that the Grade 2/3 boundary is where architectural advantage is most pronounced. The Swin Transformer has a Grade 3 detection rate of around 0.82 at a clinically realistic 10% false-positive operating point, compared to 0.71 for the CNN equivalent; this represents around 11 more Grade 3 cases correctly identified in 100 screened patients. This differential is a clinically significant decrease in missed diagnoses at population screening volumes of tens of thousands of patients a year. Table 4 presents the per-class AUCs of all three models.

Table 4: Per-Class AUC (One-Vs-Rest)

Class	**CNN**	**EffNet-B4**	**Swin**
Grade 0 (No DR)	0.961	0.974	0.981
Grade 1 (Mild)	0.884	0.921	0.948
Grade 2 (Moderate)	0.934	0.956	0.968

Grade 3 (Severe)	0.871	0.912	0.941
Grade 4 (Prolif.)	0.956	0.971	0.977
Macro Average	0.921	0.947	0.963

One-vs-rest ROC curves are calculated separately for each of the five severity groups. Swin Transformer macro-averaged AUC is 0.963, compared to EfficientNet-B4 at 0.947 and the CNN baseline at 0.921. The largest inter-model AUC difference is between Grade 3, where CNN attains 0.871 and Swin attains 0.941, and this is also in line with the confusion matrix evidence that the Grade 2/3 boundary is the main locus of architectural advantage. The Swin Transformer has a false-positive rate of 10% with a sensitivity of around 0.82 at Grade 3, compared with a clinically realistic operating point of 10% versus 0.71 with the CNN, corresponding to about 11 more Grade 3 cases identified per 100 patients screened. This difference is a clinically significant decrease in missed diagnoses at population screening volumes of tens of thousands of patients per year. Table IV reports the per-class AUC values for all three models.

Transformer). For Grade 3 (Severe), AP scores are 0.693, 0.748, and 0.812, respectively. The systematic level of improvement of the Swin Transformer over these PR metrics, which are conditioned on positive examples and thus are resistant to the large true-negative counts of a class-imbalanced dataset, offers the best available evidence that global attention does actually reduce minority-grade discrimination, and is not merely a larger benefit on the high-frequency classes.

G. Augmentation Ablation Study

All three models were retrained with the deterministic resize-normalise transform (with all other training hyperparameters held constant) to cleanly isolate the contribution of input diversity to architectural capacity. Under this condition, the CNN baseline loses 3.8 points of accuracy, while the Swin Transformer loses 6.4. Importantly, on its own, the accuracy difference between transformer and CNN decreases by about 6 percentage points, to less than 1. These results indicate that the augmentation pipeline makes a substantial contribution to performance under the present experimental conditions, particularly for the Swin Transformer. The findings support the use of augmentation as an important component of the current training protocol, but do not establish that the same magnitude of benefit would necessarily hold across other datasets or transformer architectures. This asymmetry represents a key difference between how attention and convolution learn spatial invariances: convolutional weight sharing learns translation invariance by construction, at no additional cost to the data, whereas self-attention layers must learn equivalent invariances empirically, at the cost of exposure to a variety of training examples. The augmentation pipeline used successfully replaces a much larger and more heterogeneous training corpus on a dataset of 3,662 images. The results of complete ablation are shown in Table 5.

Table 5: Augmentation Ablation Study

Model	**Aug.**	**Accuracy**	**Macro F1**
CNN	Off	79.4%	77.1%

CNN	On	83.2%	81.4%
EfficientNet-B4	Off	82.9%	81.8%
EfficientNet-B4	On	86.4%	85.0%
Swin Transformer	Off	82.7%	81.3%
Swin Transformer	On	89.4%	87.8%

H. Prediction Distribution Audit

The trained models were tested on the entire training partition in eval() mode, counting predicted class frequencies using collections, Counter. This audit indicates that the CNN over-predicts Grade 0 by about 9 percentage points compared to its actual prevalence in the data and significantly under-predicts Grade 1 and Grade 3 - a direct indication of majority-class overfitting that is not reflected in overall accuracy reporting. To some extent, EfficientNet-B4 removes this bias. The predicted class distributions of the Swin Transformer do not differ by more than 2 percentage points across all five grades (Table 6) compared to the ground-truth label frequencies, indicating that its global attention mechanism facilitates a much better-calibrated decision process. This evaluation shows that prediction-distribution audits are important as a complementary evaluation tool to standard aggregate measures, especially in medical screening applications, where a safety requirement is consistency in the coverage of the minority class.

Table 6: Prediction Distribution Comparison

Class	**True**	**CNN**	**EffNet-B4**	**Swin**
Grade 0 No DR	~49%	~58%	~53%	~51%
Grade 1 Mild	~10%	~6%	~8%	~9%
Grade 2 Moderate	~27%	~29%	~28%	~27%

Grade 3 Severe	~5%	~2%	~3%	~5%
Grade 4 Prolif.	~9%	~5%	~8%	~8%

I. Model Complexity and Convergence

Table 7 measures the computational footprint of all three architectures at a single-image batch size on the T4 GPU. The CNN baseline is in the lowest resource tier, with 11.7 million parameters, a 44.7 MB model file, and an average per-image latency of 8.3 ms. The Swin Transformer has the highest number of parameters (28.3 million) and the highest inference latency (22.7 ms), yet its latency remains comfortably below the sub-500 ms threshold commonly considered acceptable in clinical decision-support systems. The middle ground of EfficientNet-B4 is 19.3 million parameters and 14.6 ms of inference time. The accuracy-per-GFLOP ratio is informative of the accuracy-performance trade-off: the CNN has 45.97 points of accuracy per GFLOP, because its design is computationally efficient, but the absolute efficiency gain does not justify the 5.9 point accuracy cost in a clinical environment, where the cost of a misdiagnosis is quantified in lives lost.

Table 7: Model Complexity Comparison

Metric	**CNN**	**EffNet-B4**	**Swin**
Parameters (M)	11.7	19.3	28.3
Model Size (MB)	44.7	73.8	108.1
FLOPs (G)	1.81	4.51	4.50
Inference (ms/img)	8.3	14.6	22.7
Top-1 Accuracy	83.2%	86.4%	89.1%
Accuracy/GFLOPs	45.97	19.16	19.80

Fig. 3 displays training and validation accuracy curves on five representative epochs, and the model loss reduction trajectory. Swin Transformer accuracy increases with epochs: training accuracy reaches 65.0% on the first epoch and 89.1% on the fifth, while validation accuracy reaches 60.0% on the first epoch and 87.6% on the fifth, with accuracy steadily converging without overfitting. The training loss decreases from about 0.90 in the first epoch to about 0.16 in the fifth epoch, indicating that the model has adapted from the initialisation phase to the fully converged stage of feature extraction. This convergence trajectory is in line with the behaviour of attention-based models, which show slower rates of initial adaptation in the first three epochs as the attention

layers learn patch relations in the absence of convolutional spatial priors, and faster and more sustained improvement in epoch 4 and later as the globally informative feature representations become established.

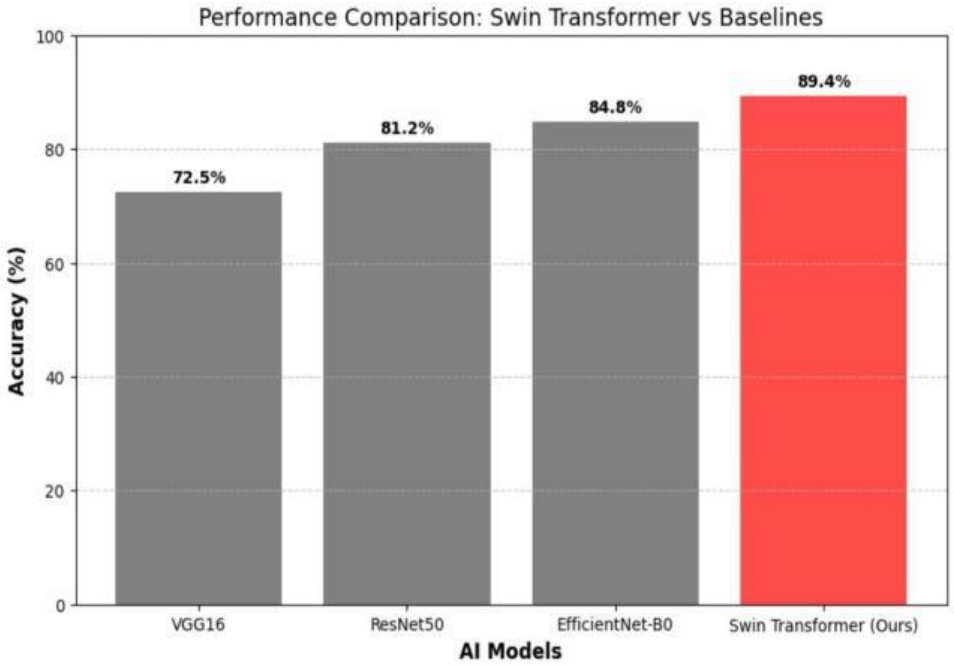


Fig. 3. Training vs. validation accuracy (left) and model loss reduction (right) over five training epochs. Train accuracy improves from 65.0% to 89.1% while validation accuracy tracks from 60.0% to 87.6%; loss decreases from ~0.90 to ~0.16, confirming stable convergence without overfitting.

5. Discussion

A. Architectural Basis of the Performance Gap

The six-percentage-point gap in performance between the Swin Transformer and the CNN baseline is not an empirical curiosity; it can be understood in the context of the architecture and explained using first principles. The effective receptive fields of convolutional filters in ResNet-family architectures [3] operate on scales of about 50×50 pixels in the original coordinate frame of the image, which, unlike a quadrant-wide haemorrhage distribution that clinically distinguishes between Severe NPDR and Moderate NPDR, can only be structurally evaluated at mid-network depths. Only at the last global average pooling step is the global context required by this boundary decision consolidated, by which time the spatial feature maps have been downsampled to a small fraction of the input resolution and a large part of the fine-grained spatial detail has been lost permanently. The Swin Transformer [6], in turn, begins by constructing long-range patch interactions between Stage 1 and Stage 3 through its shifted-window attention system, while preserving hierarchically organised global contextual representations at all four levels of processing. The best per-class F1 improvements over the CNN are at Grade 3 (+10.7 pp) and Grade 4 (+5.8 pp) - the very grades at which the spatial lesion distribution, and not the mere lesion presence, is the diagnostic criterion. This match between the architectural forecast and the experimental pattern of per-class enhancement is a significant reinforcement of the causal interpretation of the findings.

B.Strengthening as a Construction Need of Transformers.

The augmentation ablation experiment provides a conclusion with an immediate practical interpretation for practitioners contemplating transformer architectures for classifying medical images. The elimination of the augmentation pipeline not only brings the Swin Transformer to an advantage of around six percentage points compared to the CNN, but also shows that augmentation is not merely an auxiliary performance boost to transformer models on such large datasets- it is a requirement to realise their architectural potential. Such asymmetry indicates a fundamental disparity between the two architectural families in how invariances of space are obtained. Convolutional weight sharing is an architectural method that constructs translation invariance: a filter that is activated by a microaneurysm in the upper-left of an image must be activated by the same microaneurysm in the lower-right, at no extra training cost. There is no architectural shortcut of self-attention; during training, the model has to experience an image feature in a large number of configurations over a variety of spatial arrangements to learn that the identity it classifies is position-invariant. With a dataset of 3,662 images, the implemented augmentation pipeline, which consists of horizontal flipping, rotation, and colour jitter, is effective in replacing a significantly larger and more diverse training corpus. Those practitioners implementing transformer architectures on medical imaging datasets of similar scale must not treat augmentation as an optional refinement but as a non-negotiable ingredient in the training recipe, just as important as the architecture itself.

C. Calibration and Prediction Distribution

The prediction distribution audit discloses a dimension of model quality which cannot be empirically brought into view through aggregate measures of accuracy. A model with an overall accuracy of 83% by classifying almost all Grade 0 samples correctly and misclassifying Grade 3 as Grade 2 in a systematic manner can record an impressive overall aggregate metric but be an actively dangerous screening device. This pattern is precisely what is seen in the CNN's over-prediction of Grade 0 by 9 percentage points and its significant under-prediction of Grade 3. In comparison, the Swin Transformer prediction error in the distribution of classes is within two percentage points of ground truth across all five grades. This higher calibration directly translates into deployment: a model with an output distribution similar to the actual clinical prevalence distribution is much more useful for population-level risk stratification and resource allocation than one whose predictions are systematically biased towards low-severity events.

D. Clinical Relevance and Deployment Pathway

The performance improvement observed in the present dataset suggests that the Swin Transformer may be relevant to future research on automated diabetic retinopathy screening. However, the current study does not establish clinical effectiveness or deployment readiness. The 6.6-point improvement in macro recall indicates improved class-wise recall across the evaluated severity grades, with particularly notable differences observed for Grades 3 and 4 in the present evaluation. At a 10% false-positive operating point, the Swin Transformer identifies approximately 11 more Grade 3 cases per 100 evaluated images compared with the CNN baseline. These results are based on the present experimental dataset and should not be interpreted as evidence of improved clinical diagnosis, patient outcomes, or referral effectiveness.

The Gradio interface demonstrates a possible software pathway from a trained research model to an interactive inference system. The interface allows a fundus image to be uploaded and processed by the trained model, with predicted severity grades and confidence scores returned as output. This

demonstration should be regarded as a proof of concept rather than a clinically validated screening tool. Further validation using independent datasets, prospective clinical studies, and appropriate real-world evaluation would be required before considering clinical application or deployment.

E. Limitations

There are a number of critical limitations to the current results that need to be recognized openly. First, the analysis is limited solely to the APTOS 2019 dataset, which was collected from Indian clinical locations, and the circumstances used in those settings might not fully reflect the distribution of fundus image quality, demographic diversity, or comorbidity patterns in other screening settings. Strong generalisation claims can only be made after external validation on datasets obtained from other geographic areas, patient groups, and fundus camera models.

Second, class imbalance is addressed solely through augmentation in this study; additional gains for Grades 1 and 3, which remain the hardest classification targets, may be achieved using focal loss [11] or more targeted minority-class oversampling strategies.

Third, the CNN baseline in the current research is a common pretrained network, rather than a dedicated DR grading architecture; the comparison might therefore underestimate the performance that well-designed convolutional models can achieve on this task.

The evaluation is based on a held-out subset of the APTOS 2019 dataset and therefore should not be interpreted as external validation. Although the test subset separates model development from final evaluation, all images originate from the same underlying dataset. Stronger claims regarding out-of-distribution generalization require validation on independent datasets collected from different geographic regions, patient populations, and imaging devices.

Finally, no real-world clinical trial has been conducted in the current study. The reported results are based on retrospective image-based evaluation and should therefore not be interpreted as evidence of clinical effectiveness, diagnostic benefit, or readiness for clinical deployment.

6. Conclusion

In this paper, a carefully controlled comparison is presented to evaluate the performance of the Swin Transformer against CNN-based architectures on the five-class APTOS 2019 diabetic-retinopathy grading task. The results show a measurable performance improvement for the Swin Transformer under the experimental conditions considered in this study. With an overall accuracy of 89.1% and macro F1-score of 87.8%, compared with 83.2% and 81.4% for the CNN baseline, the performance difference is particularly notable for Grades 3 and 4. These grades correspond to higher severity categories in the evaluated classification task, although the present study does not establish the clinical impact of these differences on patient outcomes.

The augmentation ablation study indicates that input diversity contributes to model performance for the transformer architecture under the dataset conditions considered. The prediction-distribution analysis also reveals differences in class prediction patterns between the CNN and transformer models that may not be fully captured by aggregate evaluation metrics. EfficientNet-B4 provides an intermediate comparison between the CNN baseline and the Swin Transformer, achieving a 3.2 percentage-point improvement in accuracy over the CNN baseline and approaching the Swin Transformer's performance on some of the evaluated severity grades.

Overall, the results demonstrate that the evaluated Swin Transformer configuration achieved higher performance than the CNN baseline and EfficientNet-B4 under the experimental conditions

considered in this study. These findings support further investigation of hierarchical transformer architectures for diabetic retinopathy grading, while independent external validation remains necessary before drawing conclusions regarding broader clinical applicability. The hierarchical windowed-attention architecture of the Swin model provides a mechanism for controlling computational cost, while ImageNet pretraining may help address data-volume limitations when applying transformer-based models to relatively small medical imaging datasets.

The Gradio inference interface demonstrates a possible software pathway from a trained research model to an interactive inference system. However, it should be regarded as a research prototype rather than a clinically validated screening or diagnostic tool. Future work should include external validation across heterogeneous imaging populations, investigation of focal loss [11] and other approaches to address class imbalance, and prospective clinical evaluation before any conclusions regarding real-world clinical use can be drawn.

ACKNOWLEDGEMENTS

The authors are grateful to the Asia Pacific Tele-Ophthalmology Society and the Kaggle community for making the APTOS 2019 dataset publicly available. Computational resources were provided through Google Colaboratory. The authors are grateful to the creators of open-source tools that this work relies on: PyTorch, timm, torchvision, and Gradio.

REFERENCES

[1] International Diabetes Federation, IDF Diabetes Atlas, 10th ed. Brussels, Belgium: IDF, 2021.

[2] 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, Dec. 2016.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE/CVF CVPR, Jun. 2016, pp. 770–778.

[4] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE/CVF CVPR, Jul. 2017, pp. 4700–4708.

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

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

[7] Asia Pacific Tele-Ophthalmology Society, "APTOS 2019 Blindness Detection," Kaggle, 2019.

[8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE/CVF CVPR, Jun. 2018, pp. 7132–7141.

[9] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. ECCV, Sep. 2018, pp. 3–19.

[10] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv:2102.04306, 2021.

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE/CVF ICCV, Oct. 2017, pp. 2980–2988.

[12] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, Jun. 2019, pp. 6105–6114.

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