MSM2: Multi Scale and Multi Model based Diabetic Retinopathy detection with TTA

Rounak Gera[1], Himanshu Verma[2]
{rounak.gera@s.amity.edu[1], himanshuv@nitj.ac.in[2]}

ASET, Amity University Punjab, Mohali, India[1], DoCSE, Dr. BR Ambedkar NIT Jalandhar, India[2]

Abstract. Diabetic Retinopathy has become a significant problem across the globe, causing blindness, due to which its identification has gained prime importance so that it can be cured at early stages. Many researchers have proposed machine learning and deep learning-based methods, but there is still room for improvement. So, we propose a method for its detection which uses multi-scale ensembling, global weighted average pooling (GWAP) with Dihedral TTA(Test Time Augment), which helps to improve the results. Additionally, we have employed a range of augmentations to enhance generalization and utilized oversampling to address class imbalance on the APTOS 2019 dataset. Our approach demonstrates significant improvement across all evaluation metrics. Using Dihedral TTA helped boost precision, recall, and F1 score compared to normal inference without dihedral TTA, resulting in more stable and reliable DR severity predictions.

Keywords: AI-Healthcare, DR, Diabetic Retinopathy, Healthcare AI

1 Introduction

Diabetic Retinopathy (DR) is a disease that progresses through stages and ultimately leads to irreversible loss of eyesight. It damages retinal blood vessels and can be classified into five groups: No DR, mild DR, moderate DR, severe DR, and proliferative DR. These gradings are challenging to perform. The traditional method of manual identification is time-consuming, costly, and prone to human errors. In today's scenario, where the number of DR cases is increasing rapidly, there is a need for automated systems to identify and grade DR, allowing treatment to be initiated at early stages.

Clinically, the stages of diabetic retinopathy represent critical junctures in disease propagation. Stage 0(No DR) - the person is perfectly fine, Stage 1(mild DR) - Presesnce of microaneurysms refering begining of retinal damage, Stage 2(moderate DR) - Increased hemorrhages and venous beading, requiring proper attention, Stage 3(severe DR) - Extensive ischemia, with high risk of developing proliferative disease and Stage 4(Proliferative DR) - Growth of abnormal blood vessels, leading to threat to vision and requires urgent treatment. These stages are important for guiding

IICTDS 2025, December 19-20, Chandigarh, India
Copyright © 2026 EAI
DOI 10.4108/eai.19-12-2025.2362709

clinical decision making about diabetic retinopathy, also helping in making an accurate and robust automated DR grading system.

For the purpose of identifying diabetic retinopathy and grading it at early stages, recent improvements in deep learning have proven quite robust; however, there is still room for improvement, as numerous challenges exist within the data itself, including highly imbalanced datasets, low quality, and inconsistent images.

Prior DR studies mostly used single CNN backbones or apply TTA and ensembling seprately. However, their combined effect specially on such a small, noisy, and imbalanced datasets such as APTOS2019 remains understudied. Motivated by this gap, our paper develops a multi-scale ensemble architecture with GWAP and dihedral TTA, and evaluates it properly across multiple baselines, with ablation and external validation. Our contribution lies in the pipeline design, analysis, and generalization study, rather than in any single component used independently.

To address these challenges, we made several decisions and choices regarding techniques and resources. For the dataset, we chose the APTOS 2019 dataset, which had an appropriate number of images, although it was also imbalanced; however, it was labelled compared to some other unlabelled datasets. Then, for data preprocessing, we first removed the extra black regions from the picture, which did not aid in prediction. Then, we randomly oversampled to equalize class counts, and we applied nine different augmentations to enhance the model's learning. Accompanying the high number of augmentations, we used a multiscale ensemble for the model, wherein three different models processed the same image at different resolutions, and the predictions were then combined to produce better results. For pooling, we used GWAP[1], which effectively captured the global meaning of the image and performed much better than traditional average and global pooling. Apart from training, the final element, i.e., inference and evaluation, for which we used Dihedral TTA, wherein the same image was passed to the model but with different augmentations, which significantly improved the model's performance on all metrics. Then, for comparison with other models, we used the F1 score, AUC [2], and Accuracy.

Diabetic Retinopathy (DR) is a widespread disease that leads to loss of vision, making early detection and grading of DR crucial. However, manual detection is time-consuming and expensive. The increasing number of cases of diabetic retinopathy necessitates the automation of diabetic retinopathy detection and grading. For this, many researchers have developed deep learning-based solutions, but these solutions face challenges such as imbalanced datasets, variations in images (including brightness and lighting), and noise, making it difficult for models to generalize and provide robust predictions. This usually leads to overfitting. Motivated by these issues, we developed an approach that accurately generalizes the problem and provides robust predictions, solving all the challenges using techniques such as data augmentation, a multi-scale ensemble model, GWAP[1], and TTA.

Our primary objective is to develop a deep learning model that can accurately predict DR gradings, overcoming challenges such as imbalanced datasets, overfitting, noisy images, poor generalization, and others. The Specific goals of our study are as follows:

- Making a robust deep learning based model to classify the DR grades, for this, we used a multi-scale ensemble with three different models and GWAP[1] as a pooling layer for all three models.

- Addressing the class imbalance as class 0, i.e., no DR, is the most prevalent class, accounting for approximately 50
- Making the final prediction more robust and accurate, for which we went for Dihedral TTA(Test Time Augment), which ensures that the predictions are not biased and also confirms the prediction in various augments to choose the most accurate and prevalent prediction.

Our contribution through this research can be concluded to

- Implementation of Multi-scale and Multi-model Ensemble with GWAP[1](Global Weighted Average Pooling) for diabetic retinopathy.
- Use of Dihedral TTA for robust and accurate predictions

The paper is structurally organized into the following sections: **Section 1** contains an introduction about diabetic retinopathy, the objective of our study, our contribution through this paper, and the problem statement. **Section 2** contains a literature survey where the analysis of recent papers related to diabetic retinopathy is presented. **Section 3** contains the main problem statement of our paper to define our purpose clearly. **Section 4** explains the proposed methodology, including the dataset, pre-processing steps, model's architecture, and inference techniques. **Section 5** contains details about Experimentation done to arrive at the final methodology and best techniques and hyperparameters, along with the model's progress across metrics at different levels of training, i.e., 5, 10, and 15 epochs. **Section 6** shows a comparison with other methods, where our method outperforms all on all metrics. **Section 7** contains all the important points that need to be properly discussed. **Section** 8 concludes with discussion of future scope and limitations.

1.1 Key Contributions

Although ensembling, GWAP, oversampling, and test time augmentation (TTA) have been explored individually in prior research papers, their combined use for DR grading specifically on such a small and highly imbalanced datasets i not properly studied. Our novelty lies not in introducing a new standalone module, but in the design, integration, and analysis of a unified multi-scale ensemble pipeline.

Specifically, we ensemble EfficientNet B0, B1 and B2 to capture complementary multi-scale receptive fields, fuse them using GWAP to emphasize discriminative lesion regions, and reduce prediction variance through dihedral TTA. Unlike prior DR papers that apply single model or isolated TTA, we perform a proper evaluation across multiple baselines, provide a component wise ablation, and show external generalization on Messidor2 dataset.

2 Literature Survey

Recent studies on diabetic retinopathy grading have explored a wide range of deep learning techniques, including CNN-based architectures, autoencoders, ensemble models and multi stage

Table 1: Literature survey of recent papers on Diabetic Retinopathy

Ref.	Objective	Contribution	Strength	Limitation
[3]	DR classification assisted with GAN [4] and autoencoder-based synthetic data	Used GAN[4] and autoencoders to synthesize new data and handle class imbalance and performed DR classification using EfficientNetB0	Synthetic dataset generated using GAN[4] and variational autoencoders	Lack of Interpretability
[5]	Detection of the DR from a colored retinal image	Utilized a Dual Features Integrated Classification (DD-FIC) model for DR detection and used Wavelet integrated Retinex (WIR) for image enhancement	used an attention mechanism to detect feature relevance	synthetic or real-world clinical deployment not explored.
[6]	DR Detection with uncertainty estimation using Bayesian approximation techniques	Used DenseNet-121 to classify and Bayesian approximation techniques to estimate the uncertainty	used Bayesian approximation techniques to know uncertainty in model's predictions	computationally expensive
[7]	DR detection and evaluation of accuracy of AI in detecting DR along with analyzing benefit of using SMOTE or CLAHE	Used SMOTE[8] to do oversampling for minority classes, used CLAHE[9] to improve quality of images	Use of SMOTE[8] and CLAHE[9]	They did not specified the exact model's architecture
[10]	Making CNN based network for lesion segmentation of DR	Made Lesion Aware Network(LANet) which uses LAM(Lesion Aware Module) to capture lesions and Feature Preserve Module(FPM) to help in feature fusion	use of LANet, LAM and FPM	Limited detection of tiny lesions and bad regions
[11]	DR detection and grading along with tackling the challenge of data	used pretrained ResNet50 and EfficientNetB0[12] and then fine tuned them as a dual branch model	used a Dual branch model	misclassifications in moderate cases
[13]	DR grading using a combination of classification, ordinal regression, and regression	used combination of classification, ordinal regression, and regression heads to produce the final prediction and multi-stage training process	Use of multi-stage approach to handle differently labelled datasets	Computationally heavy
[14]	develop a fast, automated machine-learning-based DR classification	Dividing the image into patches to extract features and then further process them to obtain results	Patch-based horizontal and vertical division to extract features properly	limited reproducibility

architectures. Most of the papers have used Eyepacs[15] and APTOS dataset, which provides 5 classes based on stages of DR.

As DR classes/labels follow proper ordering and progression, some researches have used ordinal regression instead of classic multi class classification, such models learn ordered thresholds which capture the monotonic relationship among DR grades which instead of equally penalizing all misclassification, the ordinal one penalize the near grade misclassification comparatively lesser like Tymchenko et al. [13] had combined classification along with ordinal regression and regression to demonstrating that using ordinal structure help improve grading consistency and accuracy, specifically in imbalanced datasets.

Literature survey from Table 1 shows various techniques used in different papers and analyzing their objective, contributions, strengths and limitations, with the APTOS 2019 dataset being commonly used in most of the papers.

3 Problem Statement

Diabetic Retinopathy is the primary cause of vision loss in people of all ages. Making its early detection necessary for correcting it. Manual diagnosis of the retinal images is quite time-consuming and costly, making automatic detection of diabetic retinopathy important. Although much research has been done on this topic, challenges such as imbalanced datasets, lack of consistency across images, and absence of any major differentiating factor in images lead to difficulty in generalization for models, low scores on metrics, and poor predictions. Our study aims to develop a robust deep learning-based model for the early detection and classification of diabetic retinopathy grades. The proposed model aims to improve the accuracy of predictions, especially for less-represented images (i.e., minority classes), by using random oversampling to equalize the dataset distribution and employing other techniques, such as GWAP[1] and Dihedral TTA, to enhance the final prediction results.

4 Proposed Methodology

Recent advances have shown a significant improvement in deep learning-based diabetic retinopathy detection using various techniques, such as SMOTE[8], CLAHE[9], and LANet, which have led to an increase in accuracy, precision, and other metrics. However, there is still room for improvement. Our approach uses a combination of GWAP[1], careful data processing, augmentations, multi-scale ensembling, and dihedral TTA, which has shown improved results on all metrics in experimentation. We have discussed the proposed solution in detail to provide reproducibility and understanding.

4.1 Dataset

For the dataset, we explored various datasets, including Messidor, EyePACS [15], and IDRiD, among others. However, some had a low number of images, and some were unlabelled, making the process difficult. Therefore, we chose the APTOS (Asia Pacific Tele-Ophthalmology Society)

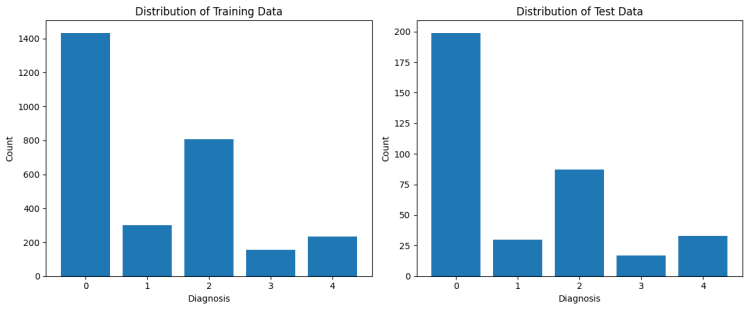


Fig. 1. Dataset distribution across train and test set, showing heavy imbalance among classes

2019 dataset, which had around 3665 images across the train, validation, and test sets, with the split samples being $2930, 366, and 366$. These image samples are collected from rural regions in India. It classifies the images into five categories, i.e., No Diabetic Retinopathy or level 0, Mild DR or level 1, Moderate DR or level 2, Severe DR or level 3, and Proliferative DR or level 4. However, the dataset is quite imbalanced, with the dominant category being No Diabetic Retinopathy or level 0. This dataset imbalance is illustrated in Figure 1, which also shows an almost equal distribution across all splits.

4.2 Data Preprocessing

For dataset preprocessing, we applied several augmentations to improve our model's learning capabilities and utilized oversampling to achieve a more balanced dataset distribution. The first step in data preprocessing is removing the black region in the image, as shown in Figure 2, which enables our model to focus only on the necessary part and exclude the unnecessary region. The second step involved oversampling where in we had done random oversampling to make the distribution of classes almost equal, so the image with level 0 were not oversampled as they were in abundance, the level 1 images were oversampled by 5 times, level 2 images were oversampled by two times, level 3 by 9 times and finally level 4 by 6 times to make the dataset balanced.

Then for augmentation for train set we used resizing, horizontal flip, vertical flip, rotation, conversion to tensor, Color jitter, gaussian bluring and erasing with the specific parameters being shown in table 2, such intensive augmentation helped to prevent our model from overfitting as generally imbalance and oversampling leads to overfitting so these prevent it and make our model generalize better. For the validation and test sets, we did not use much augmentation and resized and converted them to tensors.

4.3 Model

For the purpose of the model, we went for two major parts i.e. multiscale ensemble model and GWAP[1] pooling



Fig. 2. Difference between original image and image after removing the Noise

Table 2: Hyperparameters for all the augmentation used in train dataset

Augmentation	Parameters
Resize	224 * 224
RandomHorizontalFlip	p = 0.5
RandomVerticalFlip	p = 0.5
RandomRotation	degrees = 30
ColorJitter	brightness=0.2,contrast=0.2, saturation= 0.2, hue=0.1
GaussianBlur	kernel_size=3
RandomErasing	p=0.5

4.3.1 GWAP : Global Weighted Average Pooling

GWAP[1] helped in making the model learn as in case of diabetic retinopathy images other pooling layers like average and global pooling were not able to obtain good results so after a lot of trial and error, we came across GWAP[1] and GEM(Generalized Mean Pooling) which both provided good results so we went for GWAP[1] as it performs better in terms of feature selection. It is a type of pooling layer similar to GAP (Global Average Pooling), but it utilizes the concept of learnable weights to enhance learning and prioritize important features. It does not average all the features as GAP does, but it assigns a learnable weight to give each feature appropriate importance. Its formula is as shown below:

$$Y_c = \sum_{i=1}^{H} \sum_{j=1}^{W} W_{ij} X_{c,ij} \tag{1}$$

In this equation, $X_{c,ij}$ represents feature value at position ij and at color channel c, W_{ij} refer to learnable weights for each feature and Y_c refers to pooled feature of channel c.

4.3.2 Multi Scale Ensembling

We employed multi-scale ensembling to increase the robustness of our model and provide more accurate predictions. In this method, we used three different models and provided them with images at three different resolutions, respectively. The outputs of all models are then combined using a fully

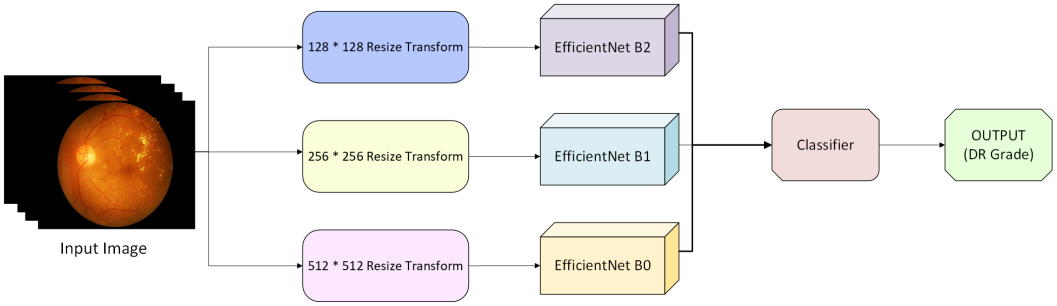


Fig. 3. Figure showing overall architecture of our multi scale ensemble model

connected layer. In our case we used three models i.e. efficientnetb0[12], efficientnetb1[12] and efficientnetb2[12], all with GWAP[1] pooling and passed images at 512*512, 256*256 and 128*128 resolution respectively. It provided a significant boost in the model's accuracy and performance.

4.4 Dihedral TTA(Test Time Augment)

For inference, we employed Dihedral TTA, which helped increase accuracy by approximately 30% across various metrics. In this, we used four augmentations, i.e., horizontal flip, vertical flip, and rotation with a 180-degree angle. Then, for each image we passed it through all augments and also original image and then for each augment and original image we passed it directly to model and passed through resized crop or zoom with scale of 0.9 and 1.0 then all the generated outputs were stored and finally for producing the final output mean or average of all values were taken to produce a final output. Table 3 shows the increase in performance after using dihedral TTA for inference.

To further analyze the impact of Test Time Augmentation (TTA), we have evaluate all used models in our approach (EfficientNet B0, B1, and B2) with and without TTA. As shown in Table 4, TTA consistently improves performance across all variants. This shows that the gains are not due to instability in a single model but a systematic effect arising from dataset characteristics such as class imbalance, limited size, and sensitivity to geometric variations.

The relatively large improvement in precision, recall, and F1-score is consistent across all backbones, suggesting that APTOS 2019's imbalance and sensitivity cause higher prediction variance in single forward passes. Dihedral TTA significantly reduces this variance by aggregating predictions across multiple transformed views. Therefore, the improvement reflects variance reduction rather than model instability.

5 Experimentation

To achieve the best combination of results, we conducted extensive experimentation. For the purpose of choosing a dataset, we reviewed several datasets, including EYEPACS[15], Messidor, and APTOS 2019. However, some datasets had a low number of images for the model to generalize

Table 3: Table demonstrating performance boost by using dihedral TTA for inference compared to without using TTA

Metric	Without Dihedral TTA	With Dihedral TTA
Accuracy	79.29%	82.67%
Precision	0.6778	0.8410
Recall	0.6403	0.8267
F1 Score	0.6419	0.8124
Kappa	0.8953	0.8810

Table 4: Comparison of Metrics for Baseline Models of our approach (With and Without TTA)

Model	Method	Accuracy	Precision	Recall	F1-Score	Kappa
EfficientNet B0	Without TTA	0.7902	0.6559	0.6328	0.6372	0.8876
	With TTA	0.8267	0.8483	0.8267	0.8308	0.8981
EfficientNet B1	Without TTA	0.7629	0.6578	0.6600	0.6361	0.8978
	With TTA	0.8352	0.8590	0.8352	0.8386	0.9043
EfficientNet B2	Without TTA	0.8065	0.7003	0.6491	0.6658	0.8871
	With TTA	**0.8551**	**0.8629**	**0.8551**	**0.8500**	0.8984

on, while others had noisy images. APTOS 2019 proved to be the most suitable. For the pooling layer, average pooling was unable to capture the global context, as it treats each region equally. In contrast, global pooling lacks parameters to adapt, making both approaches unfit. We therefore opted for GWAP[1], which utilizes weights to assign different values to various features; it is also trainable and features an attention mechanism to enhance performance.

For augmentations, we opted for those that provided greater variation in images to improve generalization. For image resolutions in the multi-scale ensemble, we previously used 1024*1024, 512*512, and 256*256, but due to the large image size, this was computationally expensive. Therefore, we opted for a combination of 512x512, 256x256, and 128x128 sizes. For the choice of models for the multi-scale ensemble, since our primary objective was to capture features from different-resolution images, we did not spend much time and opted for EfficientNetB0[12] for 512x512 images, EfficientNetB1[12] for 256x256 images, and EfficientNetB2[12] for 128x128 images. Note that all these models had a GWAP[1] pooling layer. Table 5 shows the hyperparameters used.

Table 6 shows the transformation of different metrics over time while training.

Table 5: Table with values for all the Hyper-parameters used in our approach

Hyperparameter	value
Epochs	15
Learning Rate	0.0001
Black region threshold	10
Betas	(0.9,0.999)
eps	1e-08
train test split	80 : 20

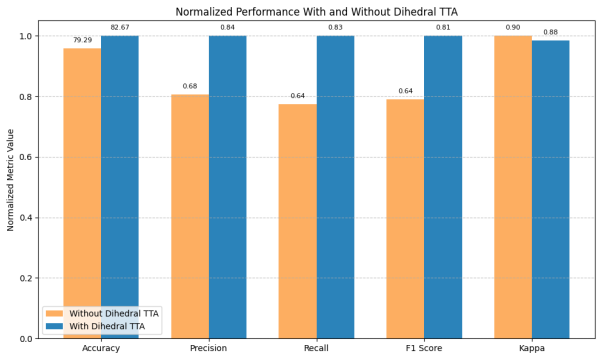


Fig. 4. Bar chart showing performance difference of inference with and without Dihedral TTA

Table 6: Metrics performance over time without dihedral TTA

	Train					Test				
Epochs	Accuracy	Precision	Recall	F1 Score	Kappa	Accuracy	Precision	Recall	F1 Score	Kappa
5	78.29	0.7859	0.7863	0.7859	0.8758	80.65	0.6986	0.6800	0.6816	0.9071
10	83.64	0.8593	0.8430	0.8729	0.9031	81.56	0.6530	0.6652	0.6761	0.9120
15	88.31%	0.8849	0.8860	0.8853	0.9319	79.29	0.6778	0.6403	0.6419	0.8953

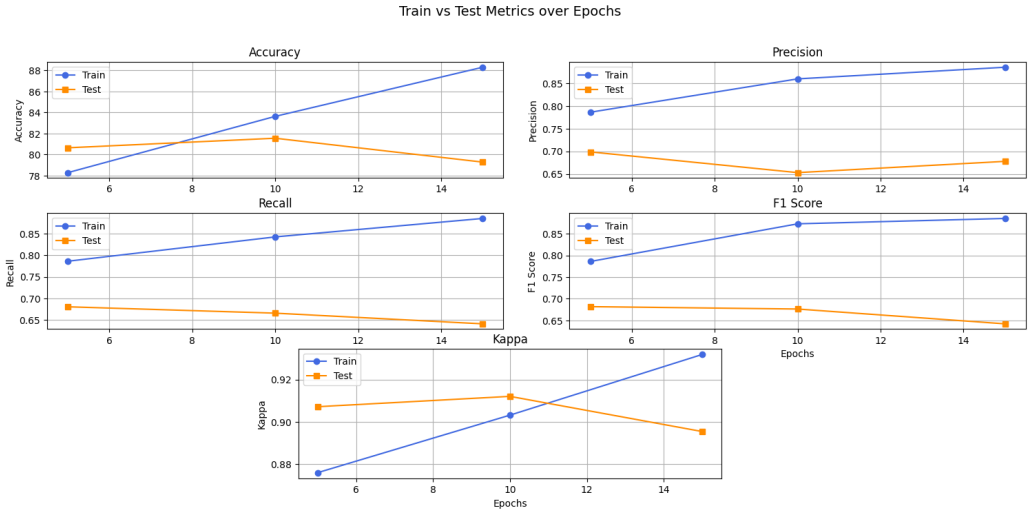


Fig. 5. Figure showing progress on metrics on different epoch levels

Table 7: Performance Comparison between our approach and existing methods from the literature.

Model	Accuracy	F1 Score	AUC[2]
ERM[16]	59.5%	41.8	78.7
Fishr[17]	63.4%	44.2	79.8
DRGen[18]	65.1%	45.3	80
Mixup[19]	51.8%	22.6	78.9
GDRNet[20]	54.8%	39.1	76.6
Ours	82.67%	81.24	85.93

Table 8: Ablation study showing the incremental contribution of each component in the proposed system.

Configuration	**GWAP**	**Oversample**	**Ensemble**	**TTA**	**Accuracy**	**F1-Score**
A1. B0 Only	✗	✗	✗	✗	78.73	58.04
A2. B0 + GWAP	✓	✗	✗	✗	79.56	62.40
A3. B0 + GWAP + Oversampling	✓	✓	✗	✗	79.02	63.72
A4. Ensemble (B0+B1+B2) + GWAP + Oversample	✓	✓	✓	✗	79.29	64.19
A5. **Full Model (Ensemble + GWAP + Oversample + TTA)**	✓	✓	✓	✓	**82.67**	**81.24**

6 Results

For comparing the contribution of TTA, we perform inference with and without Dihedral TTA and results are computed on test set using 80/10/10 split. Table 3 shows dihedral TTA produces consistent improvements across accuracy, precision, recall, and F1-score. However, Cohen's Kappa shows a slight decrease, likely due to the label distribution realignment caused by TTA.

For comparison, we evaluated our model's performance in terms of accuracy, F1 score, and AUC [2], and it shows strong performance compared to existing methods evaluated on APTOS 2019. As shown in Table 7, where we compared our model with other models, TTA has helped us achieve better results(All experiments were repeated three times, and mean values are reported to reduce variance.)

Regarding handling class imbalance, SMOTE[8] can be employed, as random oversampling does not provide sufficient variation, which is then compensated for through intensive augmentations. Additionally, from Table 6, we can easily observe that the model exhibits some degree of overfitting, as it performs better on the training set.

6.1 Ablation Study

To quantify the contribution of each component of our framework, we perform a stepwise ablation starting from a single EfficientNet model and progressively add GWAP, oversampling, ensembling, and lastly dihedral TTA. Table 8 summarizes the results.

GWAP improves representation quality by helping in combining global features, oversampling helps improve the class imbalance, and ensembling efficient net B0, B1 and B2 provides complementary multi-scale information. Finally, dihedral TTA reduces prediction variance on the small APTOS 2019 dataset, yielding the best overall performance. The although small but progressive improvements confirm that each component contributes meaningfully to the final model.

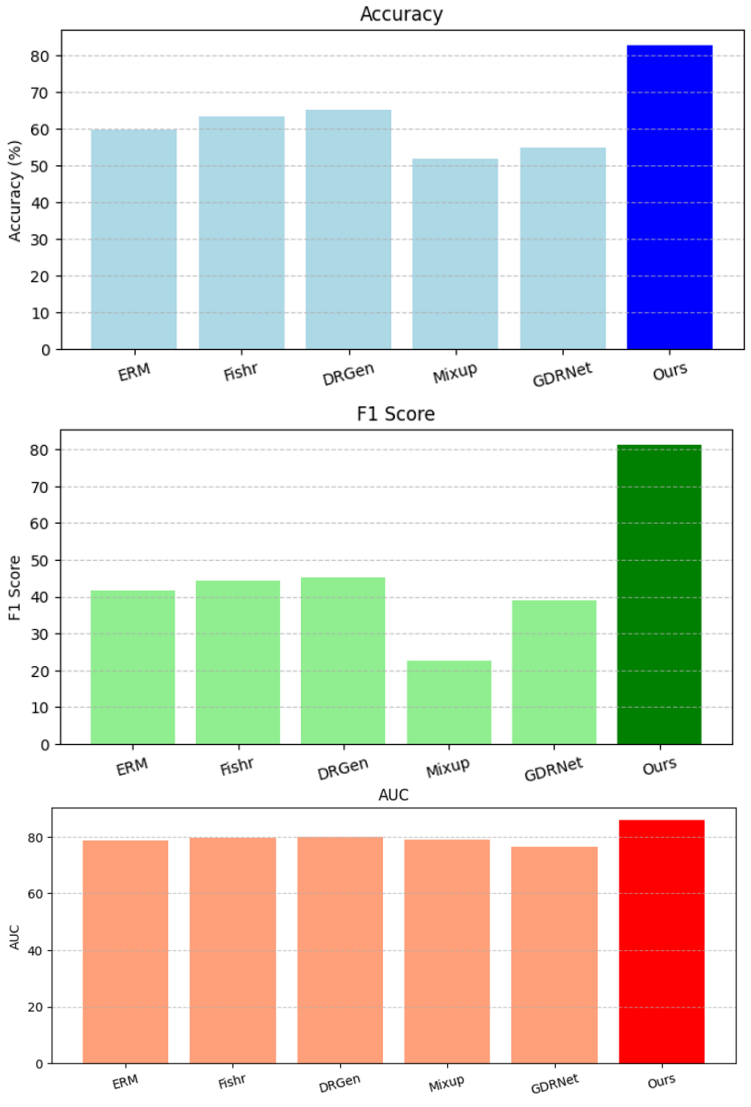


Fig. 6. Graphical comparison between proposed and existing DR models

Table 9: Comparison of model performance on APTOS2019 (in-distribution) and Messidor2 (external validation).

Dataset	Accuracy	Precision	Recall	F1-Score	Kappa
APTOS2019 (Ours)	0.8267	0.8410	0.8267	0.8124	0.8981
Messidor2 (Ours)	0.7188	0.7608	0.7188	0.6580	0.6887

6.2 External Data Validation

To assess the learning ability of the proposed model beyond APTOS 2019, we evaluate it on the Messidor 2 dataset without any retraining or fine-tuning. Messidor 2 is a widely-used benchmark for diabetic retinopathy containing images collected in different conditions. As shown in Table 9, the model maintains strong performance on this external dataset, achieving 71.88% accuracy, 0.76 precision, and 0.66 F1-score. These results demonstrate that the model generalizes reasonably well despite dataset shift.

While performance is lower on Messidor 2 than on APTOS 2019, this is expected due to differences in imaging conditions and disease distribution. These results shows the importance of cross dataset evaluation and indicate areas for future improvement.

7 Discussion

The improved performance of our model can be attributed to GWAP, Multi-Scale Ensemble, and TTA, primarily. Although minor changes, such as removing unnecessary black areas, augmentation, and oversampling, also contributed to the performance improvement. TTA helped to achieve a significant increase in metrics like accuracy, precision, recall and F1 Score with a minor decrease in kappa score which could be concluded from table 3 and figure 4 which is quite justified as by using TTA we are increasing the number of predictions the model make which can sometime lead to conflict in predictions ultimately lowering the kappa value.Although TTA provides great improvements, this effect is partially because of small size and high imbalance of APTOS2019 dataset, which increases the variance of single predictions. Thus, the gains from TTA should be interpreted as variance reduction rather than inherent model superiority. Also, a fundamental observation from figure 5 shows that the model is clearly overfitting even after all the augmentations, regularizations, and ensembling, which is quite justified as the imbalance of the dataset was handled using random oversampling, which tends to increase the chances of overfitting as similar images get repeated many times in the dataset. This can be improved using techniques like SMOTE, but it increases the computation cost, so we avoided it for now as the results were satisfactory.

8 Conclusion

Our work demonstrates that a multiscale ensemble with GWAP [1] can achieve competitive performance compared to other models. However, dihedral TTA can boost the performance showing improved results especially in case of precision, recall and F1 score. As we mentioned earlier,

SMOTE[8] can help generate artificial data. Additionally, we can use CLAHE[9] to optimize the use of datasets that contain noisy images. Selecting heavier and more complex models in a multiscale ensemble could also enhance performance. Although ordinal regression has been explored in prior DR research, combining an ordinal regression head with our architecture may be a potential future extension.

References

[1] Qiu S. Global Weighted Average Pooling Bridges Pixel-level Localization and Image-level Classification; 2018. Available from: https://arxiv.org/abs/1809.08264.

[2] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition. 1997;30(7):1145-59. Available from: https://www.sciencedirect.com/science/article/pii/S0031320396001422.

[3] GENCER K, GENCER G, CERAN TH, BİLİR AE, DOĞAN M. Photodiagnosis with deep learning: A GAN and autoencoder-based approach for diabetic retinopathy detection. Photodiagnosis and Photodynamic Therapy. 2025;53:104552. Available from: https://www.sciencedirect.com/science/article/pii/S157210002500081X.

[4] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al.. Generative Adversarial Networks; 2014. Available from: https://arxiv.org/abs/1406.2661.

[5] Devi TM, Karthikeyan P, Kumar BM, Manikandakumar M. Diabetic retinopathy detection via deep learning based dual features integrated classification model. Technology and Health Care. 2025;33(2):1066-80. PMID: 40105166. Available from: https://doi.org/10.1177/09287329241292939.

[6] Akram M, Adnan M, Ali SF, et al. Uncertainty-aware diabetic retinopathy detection using deep learning enhanced by Bayesian approaches. Scientific Reports. 2025;15:1342. Available from: https://doi.org/10.1038/s41598-024-84478-x.

[7] Mardianta S, Affandy, Supriyanto C, Supriyanto C, Wijaya A. Diabetic Retinopathy Detection Based on Convolutional Neural Networks with SMOTE and CLAHE Techniques Applied to Fundus Images; 2025. Available from: https://arxiv.org/abs/2504.05696.

[8] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Oversampling Technique. Journal of Artificial Intelligence Research. 2002 Jun;16:321–357. Available from: http://dx.doi.org/10.1613/jair.953.

[9] Mishra A. Contrast Limited Adaptive Histogram Equalization (CLAHE) Approach for Enhancement of the Microstructures of Friction Stir Welded Joints; 2021. Available from: https://arxiv.org/abs/2109.00886.

[10] Xia X, Zhan K, Fang Y, Jiang W, Shen F. Lesion-aware network for diabetic retinopathy diagnosis. International Journal of Imaging Systems and Technology. 2023 Jun;33(6):1914–1928. Available from: http://dx.doi.org/10.1002/ima.22933.

[11] Shakibania H, Raoufi S, Pourafkham B, Khotanlou H, Mansoorizadeh M. Dual branch deep learning network for detection and stage grading of diabetic retinopathy. Biomed-

ical Signal Processing and Control. 2024;93:106168. Available from: https://www.sciencedirect.com/science/article/pii/S174680942400226X.

[12] Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks; 2020. Available from: https://arxiv.org/abs/1905.11946.

[13] Tymchenko B, Marchenko P, Spodarets D. Deep Learning Approach to Diabetic Retinopathy Detection; 2020. Available from: https://arxiv.org/abs/2003.02261.

[14] Kobat SG, Baygin N, Yusufoglu E, Baygin M, Barua PD, Dogan S, et al. Automated Diabetic Retinopathy Detection Using Horizontal and Vertical Patch Division-Based Pre-Trained DenseNET with Digital Fundus Images. Diagnostics. 2022;12(8). Available from: https://www.mdpi.com/2075-4418/12/8/1975.

[15] Cuadros J, Bresnick G. EyePACS: An Adaptable Telemedicine System for Diabetic Retinopathy Screening. Journal of Diabetes Science and Technology. 2009;3(3):509-16. PMID: 20144289. Available from: https://doi.org/10.1177/193229680900300315.

[16] Gulrajani I, Lopez-Paz D. In search of lost domain generalization. arXiv preprint arXiv:200701434. 2020.

[17] Rame A, Dancette C, Cord M. Fishr: Invariant gradient variances for out-of-distribution generalization. In: International Conference on Machine Learning. PMLR; 2022. p. 18347-77.

[18] Atwany M, Yaqub M. DRGen: domain generalization in diabetic retinopathy classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2022. p. 635-44.

[19] Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:171009412. 2017.

[20] Che H, Cheng Y, Jin H, Chen H. Towards generalizable diabetic retinopathy grading in unseen domains. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2023. p. 430-40.