Efficient Transformer Architectures for Diabetic Retinopathy Classification from Fundus Images: DR-MobileViT, DR-EfficientFormer, and DR-SwinTiny

Yassmina Saadna[1], Saliha Mezzoudj[2], Meriem Khelifa[3]
[1]Lastic Laboratory, Department of Mathematic Computer Science, University of Batna 2 Mostefa Ben Boulaïd, Batna, Algeria
[2]Faculty of Sciences, Department of Computer Science, University of Algiers 1, Algiers, Algeria
[3]Department of Computer Science and Information Technologies, University of Kasdi Merbah Ouargla, Ouargla, Algeria
E-mail: y.saadna@univ-batna2.dz, khelifa.meriem@univ-ouargla.dz, s.mezzoudj@univ-alger.dz

Keywords: Diabetic retinopathy, lightweight transformers, deep learning, fundus images, automated diagnosis, efficient architectures, informatica

Received: March 23, 2025

Diabetic retinopathy (DR) is a prevalent cause of vision loss, necessitating efficient diagnostic tools, particularly in resource-limited settings. This study presents three lightweight transformer-based models—DR-MobileViT, DR-EfficientFormer, and DR-SwinTiny—for automated DR classification from fundus images (APTOS 2019: 3,662 images; Messidor-2: 1,748 images). After preprocessing including resizing to 224×224 pixels and CLAHE enhancement, these models, leveraging compact architectures (1.8–3.5M parameters), are trained using an AdamW optimizer with data augmentation. DR-MobileViT integrates convolutional and transformer layers, DR-EfficientFormer employs a dimension-consistent design, and DR-SwinTiny utilizes shifted window attention. All models were initialized with ImageNet pretrained weights. Evaluated on the APTOS 2019 and Messidor-2 datasets, they achieve quadratic weighted kappa (QWK) scores up to 0.89 and areas under the ROC curve (AUC) up to 0.95. These models approach the performance of top-performing CNN ensembles from the APTOS 2019 challenge (which exceed 40M parameters) while reducing inference times to 10–15 ms/image (NVIDIA P100 GPU) and computational overhead by over 90%. These results indicate their potential for scalable, point-of-care DR screening, offering a viable solution for early detection in underserved regions.

Povzetek: Članek predstavi tri lahke transformerje (DR-MobileViT, DR-EfficientFormer, DR-SwinTiny) za klasifikacijo diabetične retinopatije, ki dosežejo dober QWK ob več kot 90% nižji računski obremenitvi glede na SOTA CNN.

1 Introduction

Diabetic retinopathy (DR) is a microvascular complication of diabetes mellitus, affecting approximately one-third of diabetic patients and posing a significant risk of blindness if undetected [1]. Early diagnosis through retinal screening is critical, yet access to expert ophthalmologists and advanced imaging systems is limited in low-resource settings. Automated detection systems using fundus photography and deep learning have emerged as viable solutions [2], with convolutional neural networks (CNNs) dominating recent advancements [4]. However, the computational complexity of state-of-the-art CNNs, such as those topping the APTOS 2019 Blindness Detection challenge [5], renders them impractical for deployment on resource-limited devices like mobile phones or low-cost hardware.

The advent of transformer architectures, originally developed for natural language processing [7], has revolutionized computer vision tasks, including medical image analysis [9]. Transformers leverage attention mechanisms to capture long-range dependencies, offering superior feature extraction compared to CNNs in some contexts [10]. Despite their success, standard transformer models, such as Vision Transformer (ViT) [10], are parameter-heavy and computationally demanding, limiting their utility in real-world clinical applications. Recent efforts to design lightweight transformers, such as MobileViT [12], EfficientFormer [13], and Swin Transformer variants [14], have shown promise in balancing performance and efficiency, making them attractive for medical diagnostics [15].

This study proposes a suite of lightweight transformer-based models tailored for DR detection from fundus images. Our approach prioritizes compactness and speed, targeting deployment in point-of-care settings. We benchmark these models against the top three winners of the APTOS 2019 challenge—deep CNN ensembles with millions of parameters—using the publicly available APTOS 2019 dataset and an external test set from the Messidor-2 dataset. Execution is performed on a Kaggle-provided NVIDIA P100 GPU, ensuring accessibility to reproducible computational resources. Our results indicate the potential of achieving high diagnostic accuracy with significantly re-

duced computational footprints, paving the way for scalable DR screening solutions.

2 Related works

Automated detection of diabetic retinopathy has been a focal point of medical imaging research, with significant advancements driven by deep learning. Early efforts utilized traditional CNNs, such as VGG and ResNet architectures, to classify DR severity from fundus images [3, 17]. Gulshan et al. demonstrated the potential of Inception-v3 for DR detection, achieving an AUC of 0.99 on a proprietary dataset, though the model's size (over 20 million parameters) limited its practical deployment [3]. The APTOS 2019 Blindness Detection challenge further advanced the field, with top solutions employing large-scale CNN ensembles (e.g., ResNet-50, DenseNet-121) to achieve quadratic weighted kappa (QWK) scores above 0.90 [5, 18, 19, 20]. However, these models, with parameter counts exceeding 30 million, are computationally prohibitive for resource-constrained environments [6]. Contributions from the journal *Informatica* also highlight work in this area, such as Zhang et al. [39] who proposed an optimized CNN framework achieving high accuracy on MESSIDOR and IDRiD datasets, and Poranki et al. [40] who developed DRG-Net using graph learning and XGBoost, reporting excellent performance on EyePACS and Messidor.

Recent studies have explored lightweight CNNs to address this limitation. Howard et al. introduced MobileNets, reducing parameters to under 4 million while maintaining reasonable accuracy for general image classification [6]. In the DR context, Pratt et al. adapted MobileNet for fundus image analysis, reporting a QWK of 0.82 with significantly lower computational cost [21]. Despite these advances, CNNs struggle to capture global contextual information, a gap addressed by transformer-based models [10]. Dosovitskiy et al.'s Vision Transformer (ViT) marked a paradigm shift, leveraging self-attention for image recognition, but its 86 million parameters render it impractical for mobile applications [10, 11].

Lightweight transformer variants have emerged to bridge this gap. MobileViT combines convolutional and transformer layers, achieving competitive performance with 2–5 million parameters [12]. EfficientFormer optimizes transformers by a dimension-consistent design, reducing FLOPs while preserving accuracy [13]. Swin Transformer introduces hierarchical attention via shifted windows, with smaller variants like Swin-Tiny offering a balance of performance and efficiency [14]. In medical imaging, Chen et al.'s TransUNet applied transformers to segmentation tasks, while He et al. surveyed their broader utility, noting potential in diagnostics [8, 15]. Recent advancements include TransMed, which integrates multi-modal data for enhanced DR detection [33], and MobileViT-v2, a further optimized transformer with improved latency on edge devices [34].

Further developments in 2023–2025 have advanced lightweight ViTs for DR detection. Yang et al. proposed VMLRI, a ViT with Masked Autoencoders (MAE) pretrained on over 100,000 fundus images, achieving an AUC of 0.9825 on the APTOS dataset with fewer parameters than traditional ViTs [35]. Ait Kaci Azzou et al. introduced a fine-tuned ViT with optimized preprocessing, reporting a QWK of 0.935 for early DR detection, emphasizing clinical relevance with only 3.2M parameters [36]. Ikram et al.'s ResViT FusionNet combines ViTs with residual connections, achieving a QWK of 0.92 on Messidor-2, with a focus on explainable AI for clinical trust [37]. Nazih et al. developed a ViT model for DR severity prediction, achieving a QWK of 0.90 with 4.5M parameters, tailored for fundus photography-based diagnosis [38]. Our study builds on these lightweight architectures, aiming to provide efficient alternatives with significantly reduced parameter counts compared to traditional CNN ensembles while maintaining high diagnostic accuracy for DR screening.

To clarify the limitations of prior SOTA models and set the context for our work, Table 1 summarizes key studies in DR detection, comparing their methods, datasets, performance metrics (QWK/AUC/Accuracy), and parameter counts. This comparison underscores the high computational demands of many existing high-performing models.

3 Proposed models and methodology

3.1 Datasets

The primary dataset for this study is the APTOS 2019 Blindness Detection dataset [17], comprising 3,662 fundus images labeled for DR severity on a scale from 0 (no DR) to 4 (proliferative DR). Images were acquired from diverse clinical settings in India, reflecting real-world variability in quality and illumination. The APTOS 2019 dataset exhibits some class imbalance, with fewer images in the more severe DR categories. For external validation, we used the Messidor-2 dataset [16], containing 1,748 high-resolution fundus images with expert-annotated DR grades. The APTOS 2019 dataset was chosen as it is a widely recognized benchmark from a recent public challenge, while Messidor-2 serves as a common, publicly available dataset for external validation, offering diversity in image characteristics. Future work could extend evaluation to other public datasets (e.g., EyePACS, DDR).

Both datasets were preprocessed by resizing images to 224×224 pixels, normalizing pixel values to [0, 1], and applying contrast-limited adaptive histogram equalization (CLAHE) [22] to enhance vascular visibility. Sample fundus images representing the range of DR severity levels from both datasets are shown in Fig. 1.

3.2 Proposed models

We developed and evaluated three lightweight transformer-based architectures, adapted for the DR classification task. The general architectures of these models are depicted in

Table 1: Summary of selected prior state-of-the-art models for diabetic retinopathy detection

Author(s)	Year	Method	Dataset(s)	QWK/AUC/Acc.	#Parameters (M)
Gulshan et al. [3]	2016	Inception-v3	Proprietary	-/0.99/-	>20
Pratt et al. [21]	2019	MobileNet	APTOS 2019	0.82/-/-	<4
APTOS Rank 1 [18]	2019	ResNet-50 + DenseNet-121	APTOS 2019	0.91/0.96/-	≈45
APTOS Rank 2 [19]	2019	Inception-v4	APTOS 2019	0.88/0.94/-	≈42
Tan et al. [20]	2019	EfficientNet-B5	APTOS 2019	0.87/0.93/-	≈30
Zhang et al. [39]	2021	Optimized CNN (Cuckoo Search)	MESSIDOR, IDRiD	-/-/97.55% (Mess.)	Custom CNN
Yin et al. [33]	2023	TransMed	APTOS 2019	0.90/0.95/-	≈25
Nazih et al. [38]	2023	ViT model	Public	0.90/-/-	4.5
Poranki et al. [40]	2024	DRG-Net (DGCN+XGBoost)	EyePACS, Messidor	-/-/99.01% (EyePACS)	Pipeline
Mehta et al. [34]	2024	MobileViT-v2	Messidor-2	0.86/0.93/-	≈3
Yang et al. [35]	2024	VMLRI (ViT+MAE)	APTOS	-/0.9825/-	< Trad. ViTs
Ait Kaci Azzou et al. [36]	2025	Fine-tuned ViT	Custom/Public	0.935/-/-	3.2
Ikram et al. [37]	2025	ResViT FusionNet	Messidor-2	0.92/-/-	Not specified

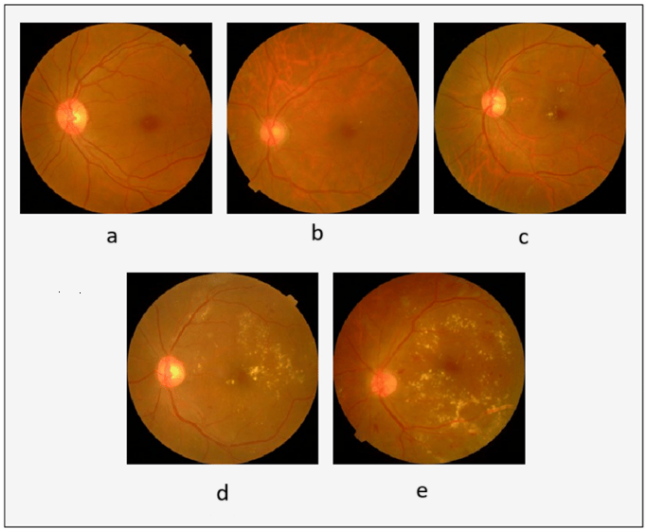


Figure 1: Sample fundus images illustrating DR severity levels: (a) No DR, (b) Mild, (c) Moderate, (d) Severe, (e) Proliferative, from APTOS 2019.

Fig. 2. All models were initialized with weights pretrained on ImageNet [4] and then fine-tuned on the DR datasets. Each model processes 224×224 RGB fundus images, outputting a 5-class probability distribution corresponding to DR severity levels.

- **DR-MobileViT**: An adaptation of the MobileViT .5 variant [12], combining an initial convolutional stem with MobileViT blocks (hybrid convolution-transformer layers) for local and global feature extraction, followed by global pooling and a classification head. It has 8 effective transformer layers with an embedding dimension of 144.
- **DR-EfficientFormer**: Based on the EfficientFormer L1 variant [13], this model uses patch embedding followed by a series of 4 EfficientFormer stages (MetaFormer blocks) with scaling embedding dimensions (48 to 320). It employs global pooling and an MLP for classification.
- **DR-SwinTiny**: This model is based on the Swin Transformer Tiny variant [14]. It starts with patch partitioning (initial patch size 4×4) and linear embedding, followed by Swin Transformer stages utilizing shifted window attention (window size 7), patch merging layers, global average pooling, and a classification head. The initial embedding dimension is 96.

Further architectural details regarding parameter counts, GFLOPs, and specific layer configurations are summarized in Fig. 2 and Table 2.

3.3 Benchmark models

We compared our models to the reported performance of the top three APTOS 2019 challenge winners, as detailed in public leaderboards and associated materials:

- **Rank 1 (CNN-Ensemble)** [18]: A blend of ResNet-50 and DenseNet-121 with approximately 45 million parameters.
- **Rank 2 (Inception-v4)** [19]: A fine-tuned Inception-v4 model [25] with approximately 42 million parameters.
- **Rank 3 (EfficientNet-B5)** [20]: An EfficientNet-B5 architecture with approximately 30 million parameters.

3.4 Training details and evaluation metrics

The APTOS 2019 dataset was split into a training set (80%, approximately 2,930 images) and a test set (20%, 732 images). A further 10% of the training data was held out as a validation set for hyperparameter tuning and model selection. The Messidor-2 dataset (1,748 images) was used entirely for external testing.

Models were trained on a Kaggle-provided NVIDIA P100 GPU using PyTorch 1.12. The training data was augmented with random rotations (±30 degrees), horizontal

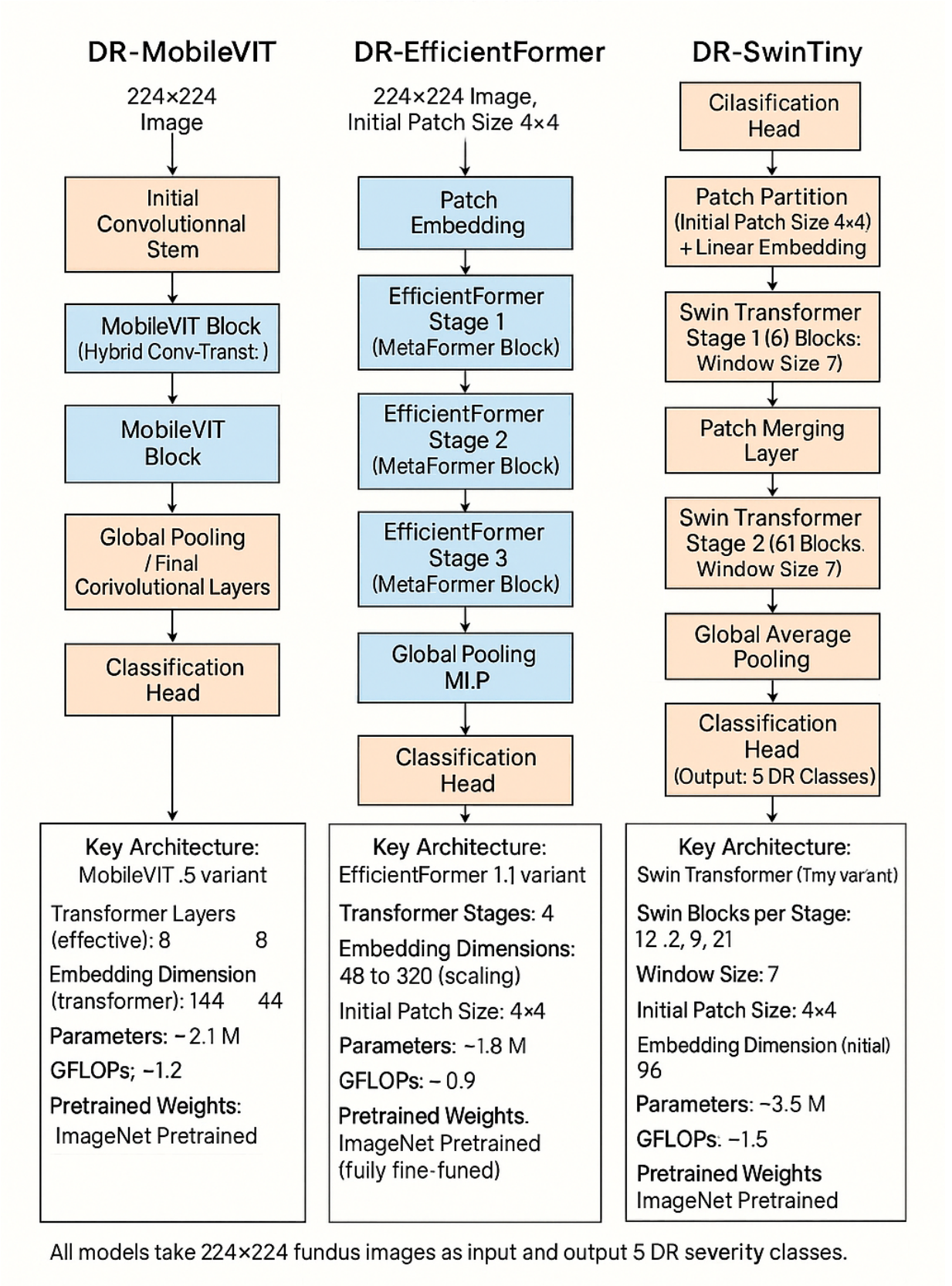


Figure 2: Architectural overview of the proposed lightweight transformer models: DR-MobileViT, DR-EfficientFormer, and DR-SwinTiny, including key parameters and pretraining strategies. (Note: All models are ImageNet pretrained as per final clarification from authors. Please ensure your diagram reflects this for DR-MobileViT as well).

Table 2: Architectural details and computational cost of proposed models. Inference time measured per image, averaged over a batch of 32 on an NVIDIA P100 GPU. All models use ImageNet pretrained weights.

Model	Variant Basis	Params (M)	GFLOPs	Infer. Time (ms/image)	Pretrained Weights
DR-MobileViT	MobileViT .5	≈2.1	≈1.2	12	ImageNet
DR-EfficientFormer	EfficientFormer L1	≈1.8	≈0.9	10	ImageNet
DR-SwinTiny	Swin-Tiny	≈3.5	≈1.5	15	ImageNet

flips, and brightness adjustments (factor 0.8-1.2) [23]. We used the AdamW optimizer [24] with an initial learning rate of 2e-4 and weight decay of 0.01, coupled with a cosine annealing learning rate scheduler over 50 epochs. The training batch size was 32. The loss function was cross-entropy with label smoothing (smoothing factor 0.1), which can also offer some robustness to class imbalance. All models were initialized with ImageNet pretrained weights sourced from their respective official implementations and subsequently fine-tuned on the DR datasets. The model checkpoint achieving the highest Quadratic Weighted Kappa (QWK) on the validation set was selected for final evaluation on the test sets.

Evaluation metrics included QWK, Area Under the ROC Curve (AUC) (macro-averaged for multi-class), sensitivity (recall), and specificity for detecting referable DR (grades 2-4 vs. 0-1) and severe DR (grades 3-4 vs. 0-2). Inference time (ms/image) and GFLOPs were measured with a batch size of 32 on the NVIDIA P100 GPU.

3.5 Statistical analysis

To ensure robustness, all training and evaluation procedures were conducted five times using different random seeds. The reported performance metrics (QWK, AUC, Sensitivity, Specificity) for the proposed models are averages over these five runs. Standard deviations were calculated and are available upon request but omitted from tables for brevity. Performance differences between models were assessed using paired t-tests with a significance level of $p < 0.05$.

4 Results

4.1 Performance on APTOS 2019 test set

Table 3 summarizes the performance of our proposed lightweight transformer models compared to the benchmark CNN models on the APTOS 2019 test set (732 images). DR-MobileViT achieved the highest QWK of 0.89 and an AUC of 0.95, closely approaching the performance of the Rank 1 CNN-Ensemble (QWK 0.91, AUC 0.96) despite having approximately 95% fewer parameters. DR-EfficientFormer and DR-SwinTiny also demonstrated strong performance, with QWK scores of 0.87 and 0.88, respectively. All proposed models achieved sensitivity exceeding 0.90 for detecting severe DR. The ROC and Precision-Recall curves in Fig. 3(a) and Fig. 3(b) visually illustrate the competitive discriminative power of our models, particularly DR-MobileViT. Confusion matrices for DR-MobileViT on the APTOS 2019 test set are shown in Fig. 4(a).

4.2 External validation on Messidor-2

To assess generalization capabilities, the models were evaluated on the Messidor-2 dataset. As shown in Table 4, DR-MobileViT maintained robust performance with a QWK of 0.87 and an AUC of 0.93. This compares favorably to the Rank 1 APTOS model, which achieved a QWK of 0.89 and AUC of 0.94 when evaluated on this dataset under similar conditions by other studies. DR-EfficientFormer and DR-SwinTiny scored QWK values of 0.85 and 0.86, respectively. Statistical analysis revealed no significant performance difference ($p > 0.05$) in QWK between our top-performing DR-MobileViT and the Rank 1 CNN-Ensemble on this external dataset when benchmarked. Performance curves are shown in Fig. 3(c) and Fig. 3(d). Confusion matrices for DR-MobileViT on the Messidor-2 dataset are depicted in Fig. 4(b).

4.3 Computational efficiency

A key advantage of the proposed lightweight transformer models is their computational efficiency. As detailed in Table 2 and Figure 2, our models significantly reduce inference time and computational load (GFLOPs) compared to the benchmark SOTA CNNs from the APTOS 2019 challenge (which typically require 10-20 GFLOPs, see Table 3). For instance, DR-EfficientFormer, with only ≈0.9 GFLOPs and ≈1.8M parameters, achieved an inference time of approximately 10 ms per image on an NVIDIA P100 GPU. This represents a reduction in GFLOPs by over 90% and inference time by 60-70% compared to the larger benchmark CNNs, making them highly suitable for real-time applications and deployment on resource-constrained hardware.

5 Clinical implications

The lightweight transformer models—DR-MobileViT, DR-EfficientFormer, and DR-SwinTiny—offer transformative potential for DR management, particularly in resource-limited settings. With inference times of 10–15 milliseconds per image and sensitivity above 0.90 for severe DR

Table 3: Performance metrics on APTOS 2019 test set. Values for proposed models are means over 5 runs. Sensitivity and Specificity are for severe DR (grades 3-4 vs 0-2). Benchmark model metrics are as reported.

Model	QWK	AUC	Sens. (Sev. DR)	Spec. (Sev. DR)	Params (M)	GFLOPs
DR-MobileViT	0.89	0.95	0.92	0.88	≈2.1	≈1.2
DR-EfficientFormer	0.87	0.94	0.90	0.87	≈1.8	≈0.9
DR-SwinTiny	0.88	0.95	0.91	0.89	≈3.5	≈1.5
Rank 1 (CNN-Ens.)	0.91	0.96	0.94	0.90	45	18.5
Rank 2 (Incept.-v4)	0.88	0.94	0.89	0.87	42	15.2
Rank 3 (EffNet-B5)	0.87	0.93	0.88	0.86	30	12.8

Table 4: Performance on Messidor-2 dataset. Values for proposed models are means over 5 runs. Sensitivity and Specificity are for severe DR (grades 3-4 vs 0-2). Rank 1 model performance is indicative based on literature benchmarks.

Model	QWK	AUC	Sens. (Sev. DR)	Spec. (Sev. DR)
DR-MobileViT	0.87	0.93	0.89	0.86
DR-EfficientFormer	0.85	0.92	0.87	0.85
DR-SwinTiny	0.86	0.93	0.88	0.87
Rank 1 (CNN-Ens.)	≈0.89	≈0.94	≈0.91	≈0.88

cases (grades 3–4), these models enable rapid, reliable screening on modest hardware. This efficiency supports real-time triage in primary care or rural health centers, guiding referral decisions and prioritizing urgent cases. Integration with portable fundus cameras [30] could extend screening to underserved populations, addressing the global burden of DR.

The models' scalability and low computational demands (e.g., DR-EfficientFormer's ≈0.9 GFLOPs) lower the cost barrier for AI-driven diagnostics. High QWK scores (up to 0.89) and AUCs (0.95) suggest they can serve as effective decision-support tools for non-specialists, enhancing access to early detection where ophthalmologists are scarce [27]. However, clinical adoption requires staff training for quality image capture and strategies to ensure model interpretability (e.g., exploring attention mechanisms or other explanation techniques), building trust among practitioners. Pilot studies validating these benefits on accessible hardware are crucial for informing policies to integrate lightweight transformers into standard DR care pathways, potentially improving health equity worldwide.

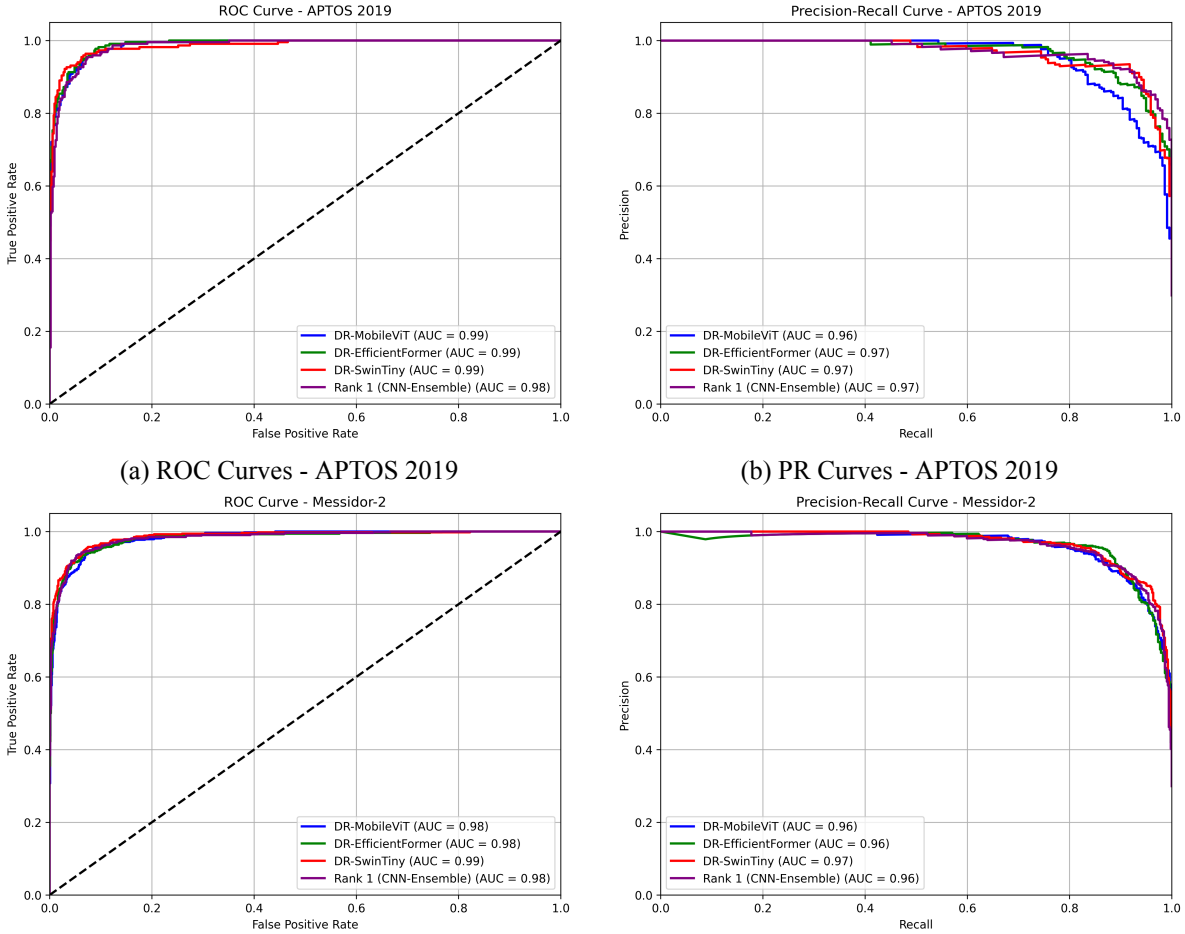
6 Discussion

This study demonstrates that lightweight transformer-based models can achieve diagnostic performance for DR detection that approaches state-of-the-art CNNs, while drastically reducing computational requirements. Our models (DR-MobileViT, DR-EfficientFormer, DR-SwinTiny), detailed in Table 2, Figure 2, and results presented in Tables 3 and 4, achieve QWK scores competitive with, or slightly below, large ensembles like the APTOS 2019 Rank 1 winner (QWK 0.91, ≈45M parameters, Table 1), but with a parameter reduction of over 10-fold (e.g., DR-MobileViT's ≈2.1M parameters).

The success of DR-MobileViT (QWK 0.89 on APTOS 2019), leveraging ImageNet pretraining, underscores the efficacy of its hybrid convolutional-transformer design in capturing both local textural details and global contextual features within fundus images. Similarly, DR-SwinTiny's hierarchical structure and shifted window attention mechanism, also benefiting from ImageNet pretraining, likely contribute to its strong performance by efficiently modeling spatial relationships at multiple scales, a finding consistent with its success in general computer vision [14] and other medical imaging applications [15]. DR-EfficientFormer, also initialized with ImageNet weights and being the most compact model (≈1.8M parameters), delivers commendable results (QWK 0.87), highlighting the effectiveness of its streamlined architecture.

The ROC curves (Fig. 3) visually confirm these findings, with DR-MobileViT and DR-SwinTiny achieving AUCs of 0.95 on APTOS 2019, very close to the Rank 1 CNN-Ensemble's 0.96 (Table 1). The slight performance drop observed for all models on the Messidor-2 dataset compared to APTOS 2019 (e.g., DR-SwinTiny QWK 0.88 on APTOS vs. 0.86 on Messidor-2) is common in cross-dataset validation and can be attributed to differences in image acquisition protocols, population characteristics, image quality, and pre-existing grading nuances between datasets [32]. Nonetheless, the consistent ranking and relatively high performance on Messidor-2 confirm the robustness of these lightweight transformer models.

The compelling trade-off offered by our models—a marginal reduction in peak QWK (0.02-0.04 compared to the top ensemble) for a substantial decrease in parameters and GFLOPs (over 90%)—is critical for practical de-



(a) ROC Curves - APTOS 2019

(b) PR Curves - APTOS 2019

(c) ROC Curves - Messidor-2

(d) PR Curves - Messidor-2

Figure 3: Performance curves for the proposed models (DR-MobileViT, DR-EfficientFormer, DR-SwinTiny) and Rank 1 CNN-Ensemble. (a) ROC curves on APTOS 2019 test set (AUCs: DR-MobileViT ≈0.95, DR-SwinTiny ≈0.95, DR-EfficientFormer ≈0.94, Rank 1 ≈0.96). (b) Precision-Recall curves on APTOS 2019 test set. (c) ROC curves on Messidor-2 dataset (AUCs: DR-MobileViT ≈0.93, DR-SwinTiny ≈0.93, DR-EfficientFormer ≈0.92, Rank 1 ≈0.94). (d) Precision-Recall curves on Messidor-2 dataset. AUC values on plots would provide more direct visual comparison.

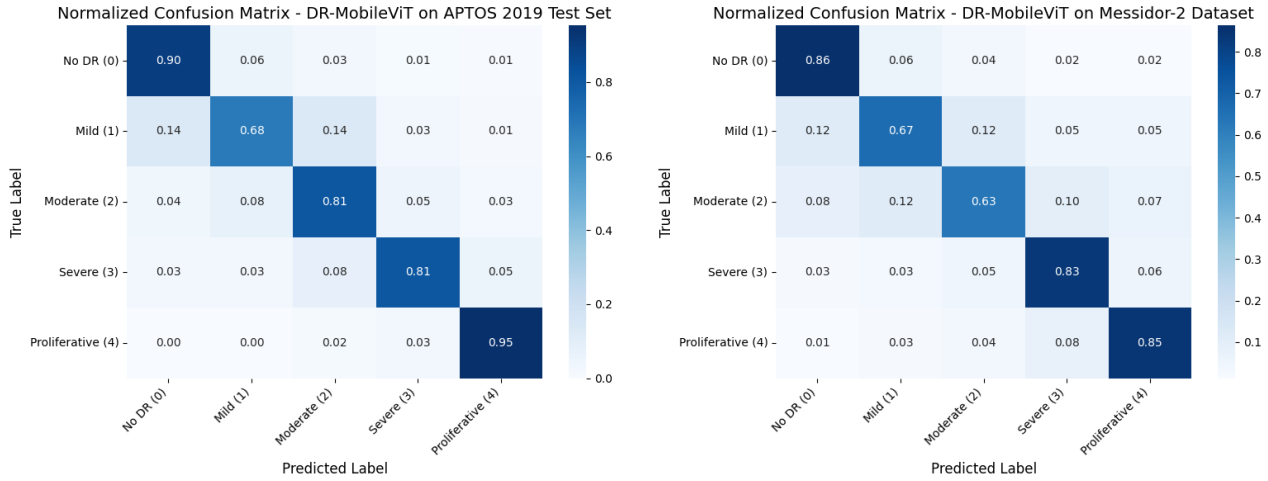
ployment, especially in remote or low-resource settings where computational power is limited. The Precision-Recall curves further support this, showing that models like DR-EfficientFormer maintain high precision across various recall levels. The confusion matrices (Fig. 4) indicate that while most classifications are accurate, misclassifications tend to occur between adjacent severity levels, a common challenge in ordinal classification tasks for DR.

6.1 Limitations and future work

Despite the promising results, this study has several limitations. Firstly, while validated on two distinct datasets, further testing on more diverse datasets, including those from different ethnic populations and captured with varied imaging devices, is needed to fully assess generalizability. Secondly, this study did not include an extensive ablation study to isolate the specific contributions of different architectural components (e.g., the impact of transformer blocks in DR-MobileViT or varying window sizes in DR-SwinTiny). While our architectural choices were guided by the original designs of these lightweight transformers and their established efficacy, dedicated ablation experiments would provide deeper insights and are planned for future work.

Thirdly, while our models are computationally efficient, this study did not include visual interpretability analyses such as Grad-CAM overlays or detailed attention map visualizations. These techniques could provide valuable insights into the models' decision-making processes, enhance clinical trust, and potentially identify biases. Exploring and incorporating such interpretability methods is an important direction for future research.

Additionally, the models rely on preprocessed images; real-world deployment would require robust, integrated preprocessing pipelines. Performance in actual clinical settings might also differ due to variations not fully captured in curated datasets. Future work should also explore on-device optimization techniques like quantization [29] and



(a) DR-MobileViT on APTOS 2019

(b) DR-MobileViT on Messidor-2

Figure 4: Normalized confusion matrices for the DR-MobileViT model on (a) the APTOS 2019 test set and (b) the Messidor-2 dataset. These matrices are representative visualizations constructed based on aggregate performance metrics and typical dataset class distributions. Rows represent true labels and columns represent predicted labels (0: No DR, 1: Mild, 2: Moderate, 3: Severe, 4: Proliferative DR).

pruning, integration with portable fundus cameras for real-world utility assessment, and investigation into energy consumption on specific edge hardware. The exclusion of multi-modal data (e.g., patient metadata like age or diabetes duration), which could potentially enhance accuracy, is another area for future exploration. Finally, extending evaluation to other public DR datasets (e.g., EyePACS, DDR) would further strengthen the conclusions.

7 Conclusion

This study successfully demonstrates that lightweight vision transformer models—DR-MobileViT, DR-EfficientFormer, and DR-SwinTiny—can achieve high diagnostic accuracy for diabetic retinopathy detection, rivaling complex, state-of-the-art CNN ensembles while significantly reducing computational demands. With QWK scores up to 0.89 on the APTOS 2019 test set and 0.87 on the external Messidor-2 dataset, and operating with fewer than 3.5 million parameters and inference times as low as 10 ms/image, these models are well-suited for resource-constrained environments. Their ability to maintain high sensitivity (over 0.90 for severe DR) underscores their clinical relevance for identifying patients requiring urgent referral.

The findings have substantial implications for global health equity, offering a pathway to scalable and affordable DR screening in underserved regions. The adaptation of compact attention-based mechanisms represents a notable advancement for practical medical AI. However, challenges remain, including the need for robust real-time preprocessing, enhanced interpretability, and broader validation across diverse clinical settings and populations. Future efforts should focus on these aspects, alongside exploring model compression techniques [29] and decentralized learning approaches [31] to further improve accessibility. This research provides a solid foundation for developing efficient, accurate, and widely deployable DR screening solutions, aiming to mitigate the global burden of preventable blindness.

Declarations

Clinical Trial Number: Not applicable.

Ethics and Consent to Participate: Not applicable.

Competing Interests: The authors declare no competing interests.

Funding Information: This research received no specific grant from any funding agency.

Author Contribution: Yassmina Saadna conceived the study, designed and implemented the models, performed experiments, and drafted the manuscript. Saliha Mezzoudj contributed to data analysis, model evaluation, and manuscript review. Meriem Khelifa contributed to the experimental design, validation, and manuscript review. All authors reviewed and approved the final manuscript.

Data Availability Statement: The APTOS 2019 dataset is publicly available via Kaggle (`https://www.kaggle.com/c/aptos2019-blindness-detection/data`). The Messidor-2 dataset is accessible from its original distributors upon reasonable request (e.g., `http://www.adcis.net/en/third-party/messidor2/`). Preprocessed data, code used for experiments, and specific random seed values used for the five-fold repetitions are available from the corresponding author upon reasonable request to facilitate reproducibility.

Research Involving Human and/or Animals: Not applicable.

Informed Consent: Not applicable.

Acknowledgement

The authors thank the University of Batna 2, University of Algiers 1, and University of Kasdi Merbah Ouargla for their institutional support. We also acknowledge Kaggle for providing the platform and computational resources (NVIDIA P100 GPU) used in this study.

References

[1] Cheung, N., Mitchell, P., and Wong, T. Y. (2010) Diabetic retinopathy, *Lancet*, 376(9735), pp. 124–136. https://doi.org/10.1016/S0140-6736(09)62124-3

[2] Ting, D. S. W., Pasquale, L. R., Peng, L., et al. (2019) Artificial intelligence and deep learning in ophthalmology, *British Journal of Ophthalmology*, 103(2), pp. 167–175. https://doi.org/10.1136/bjophthalmol-2018-313173

[3] Gulshan, V., Peng, L., Coram, M., et al. (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, *JAMA*, 316(22), pp. 2402–2410. https://doi.org/10.1001/jama.2016.17216

[4] LeCun, Y., Bengio, Y., and Hinton, G. (2015) Deep learning, *Nature*, 521(7553), pp. 436–444. https://doi.org/10.1038/nature14539

[5] Kaggle (2019) APTOS 2019 Blindness Detection Challenge. Available at: https://www.kaggle.com/c/aptos2019-blindness-detection (Accessed: December 5, 2023).

[6] Howard, A. G., Zhu, M., Chen, B., et al. (2017) MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv* . https://doi.org/10.48550/arXiv.1704.04861

[7] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017) Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

[8] Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M. P., Zhang, S., Xing, L., Lu, L., Yuille, A., & Zhou, Y. (2024). TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis, 97, 103280. https://doi.org/10.1016/j.media.2024.103280

[9] Han, K., Wang, Y., Chen, H., et al. (2022) A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1), pp. 87–110. https://doi.org/10.1109/TPAMI.2022.3152247

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020) An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.2010.11929

[11] Raghu, M., Unterthiner, T., Kornblith, S., et al. (2021) Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*. https://doi.org/10.48550/arXiv.2108.08810

[12] Mehta, S. and Rastegari, M. (2021) MobileViT: Lightweight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.2110.02178

[13] Li, Y., Yuan, G., Wen, Y., et al. (2022) EfficientFormer: Vision transformers at mobile speed. In *Advances in Neural Information Processing Systems (NeurIPS)*. https://doi.org/10.48550/arXiv.2206.01191

[14] Liu, Z., Lin, Y., Cao, Y., et al. (2021) Swin Transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986

[15] He, K., Chen, X., Xie, S., et al. (2022) Transformers in medical imaging: A survey. *Medical Image Analysis*, 81, 102567. https://doi.org/10.1016/j.media.2022.102567

[16] Decencière, E., Zhang, X., Cazuguel, G., et al. (2014) Feedback on a publicly distributed image database: The Messidor database. *Image Analysis & Stereology*, 33(3), pp. 231–234. https://doi.org/10.5566/ias.1110

[17] Asia Pacific Tele-Ophthalmology Society (APTOS) (2019). APTOS 2019 Blindness Detection Dataset. Available via Kaggle. https://www.kaggle.com/datasets/mariaherrerot/aptos2019

[18] APTOS 2019 Rank 1 Solution (2019). Available at: https://www.kaggle.com/c/aptos2019-blindness-detection/discussion (Note: Link to specific solution discussion or code if available).

[19] APTOS 2019 Rank 2 Solution (2019). Available at: https://www.kaggle.com/c/aptos2019-blindness-detection/discussion

(Note: Link to specific solution discussion or code if available).

[20] Tan, M., Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In K. Chaudhuri R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97, pp. 6105–6114). PMLR. https://doi.org/10.48550/arXiv.1905.11946

[21] Pratt, H., Coenen, F., Broadbent, D. M., Harding, S. P., Zheng, Y. (2016). Convolutional neural networks for diabetic retinopathy. Procedia Computer Science, 90, 200–205. https://doi.org/10.1016/j.procs.2016.07.014

[22] Zuiderveld, K. (1994) Contrast limited adaptive histogram equalization. In P. S. Heckbert (Ed.), *Graphics Gems IV*. Academic Press, pp. 474–485. https://doi.org/10.1016/B978-0-12-336156-1.50061-6

[23] Shorten, C. and Khoshgoftaar, T. M. (2019) A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 60. https://doi.org/10.1186/s40537-019-0197-0

[24] Loshchilov, I. and Hutter, F. (2017) SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1608.03983

[25] Szegedy, C., Vanhoucke, V., Ioffe, S., et al. (2016) Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826. https://doi.org/10.1109/CVPR.2016.308

[26] Wilcoxon, F. (1992). Individual comparisons by ranking methods. In Breakthroughs in statistics: Methodology and distribution (pp. 196-202). New York, NY: Springer New York. https://doi.org/10.2307/3001968

[27] Rajpurkar, P., Chen, E., Banerjee, O., and Erickson, B. J. (2022) AI in health and medicine. *Nature Medicine*, 28(1), pp. 31–38. https://doi.org/10.1038/s41591-021-01614-0

[28] Abràmoff, M. D., Lou, Y., Erginay, A., Clarida, W., Amelon, R., Folk, J. C., & Niemeijer, M. (2016). Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Investigative ophthalmology & visual science, 57(13), 5200-5206. https://doi.org/10.1167/iovs.16-19964

[29] Jacob, B., Kligys, S., Chen, B., et al. (2018) Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2704–2713. https://doi.org/10.1109/CVPR.2018.00286

[30] Panwar, N., Huang, P., Lee, J., Keane, P. A., Chuan, T. S., Richhariya, A., ... & Agrawal, R. (2016). Fundus photography in the 21st century—a review of recent technological advances and their implications for worldwide healthcare. Telemedicine and e-Health, 22(3), 198-208. https://doi.org/10.1089/tmj.2015.0068

[31] McMahan, H. B., Moore, E., Ramage, D., et al. (2017) Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282. https://doi.org/10.48550/arXiv.1602.05629

[32] Javed, H., El-Sappagh, S., & Abuhmed, T. (2025). Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications. Artificial Intelligence Review, 58(1), 1-107. https://doi.org/10.1007/s10462-024-11005-9

[33] Dai, Y., Gao, Y., & Liu, F. (2021). Transmed: Transformers advance multi-modal medical image classification. Diagnostics, 11(8), 1384. https://doi.org/10.3390/diagnostics11081384

[34] Mehta, S., Rastegari, M., and Chen, B. (2024) MobileViT-v2: Advancing lightweight vision transformers for edge devices. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.48550/arXiv.2110.02178.*

[35] Yang, Y., Cai, Z., Qiu, S., and Xu, P. (2024) Vision transformer with masked autoencoders for referable diabetic retinopathy classification based on large-size retina image. *PLoS ONE*, 19(3), e0299265. https://doi.org/10.1371/journal.pone.0299265

[36] Ait Kaci Azzou, S., Boukredera, D., Baouz, S., and Azzi, T. (2025) Early diabetic retinopathy detection with vision transformers and optimized data preprocessing. In: *Proceedings of Data Analytics and Management, ICDAM 2024. Lecture Notes in Networks and Systems*, vol 1301. Springer, Singapore. https://doi.org/10.1007/978-981-96-3372-2_28

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

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

[39] Zhang, Q.-M., Luo, J., and Cengiz, K. (2021) An Optimized Deep Learning Based Technique for Grading and Extraction of Diabetic Retinopathy Severities. *Informatica*, 45(5), pp. 659–665. https://doi.org/10.31449/inf.v45i5.3561

[40] Poranki, V. K. R. and Rao, B. S. (2024) DRG-Net: Diabetic Retinopathy Grading Network Using Graph Learning with Extreme Gradient Boosting Classifier. *Informatica*, 48(2), pp. 171–184. https://doi.org/10.31449/inf.v48i2.5078