PolyVision: Optimising Retinal Disease Detection Through Collaborative Neural Networks

Sultan Ahmad [1,*], Swathi Kalam [2], Eali Stephen Neal Joshua [3], Hikmat A. M. Abdeljaber [4], and Hessa Alfraihi [5]

[1] Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Alkharj, Saudi Arabia
[2] Department of Computer Science and Engineering, Vignan's Institute of Information Technology, Visakhapatnam, India
[3] Department of Computer Science and Engineering, GST, GITAM (Deemed to be University), India
[4] Department of Computer Science, Faculty of Information Technology, Applied Science Private University, Amman, Jordan
[5] Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia
Email: s.alisher@psau.edu.sa (S.A.); swathi.kalam@gmail.com (S. K.); seali@gitam.edu (E.S.N.J.); h_abdeljaber@asu.edu.jo (H.A.M.A.); haalfraihi@pnu.edu.sa (H.A.)
*Corresponding author

Abstract*—Diabetic Retinopathy (DR) remains one of the leading causes of preventable blindness worldwide, and early, reliable diagnosis is essential for reducing vision loss. Deep learning has shown promise in this domain, but single models often suffer from limited generalizability, sensitivity–specificity imbalance, and high computational demand. To address these challenges, we present PolyVision, a modular ensemble framework designed for robust and equitable DR screening. PolyVision integrates three complementary backbones—ResNet50, EfficientNet-B2, and Vision Transformer—each capturing different levels of spatial and contextual retinal features. Their predictions are combined through a dual fusion mechanism based on mean and maximum voting, which balances diagnostic sensitivity and specificity while minimizing variance across models. To further enhance robustness, the models are trained with diverse augmentation strategies, and hyperparameters are tuned for optimal performance. Evaluated on ultra-widefield fundus images, PolyVision achieved an AUC-ROC of 0.953, an AUPRC of 0.975, and an inference latency of 110 ms per image, demonstrating both high diagnostic accuracy and clinical efficiency. Beyond accuracy, the framework incorporates fairness evaluation across imaging subgroups, supporting more equitable diagnostic outcomes. Its lightweight design also facilitates deployment in resource-constrained clinical settings without compromising reliability. These results highlight the potential of ensemble learning to provide scalable, accurate, and fair DR screening. However, additional validation on multi-institutional datasets and real-world clinical environments remains necessary before broad clinical adoption.

***Keywords*—diabetic retinopathy, convolutional neural networks, vision transformer, deep learning, model fusion, fairness in AI, medical image analysis**

I. Introduction

Retinal Diseases (RD), i.e., Diabetic Retinopathy (DR), are among the top causes of visual impairment and blindness throughout the world, DR being a specific threat to the working-age population. According to the World Health Organization, DR affects more than 90 million people globally and remains a leading cause of preventable blindness [1]. In parallel with the increase in diabetes prevalence, the global burden of DR also continues to surge, thereby placing a significant strain on healthcare systems worldwide. Early detection and timely treatment have been recognized as key strategies in the attempt to reduce vision loss from DR [2].

Deep learning-based diagnostic systems have shown immense promise in augmenting clinical workflows for DR detection and monitoring. Convolutional Neural Networks (CNNs) have long dominated medical image analysis due to their capacity for efficient local feature extraction [3]. More recently, Vision Transformers (ViTs) have gained attention for their ability to capture long-range dependencies through self-attention mechanisms [4–6]. However, ViTs are computationally intensive and prone to overfitting in domains like medical imaging, where annotated data is often scarce. Conversely, CNNs—while less global in scope—offer strong performance with lower

Manuscript received July 3, 2025; revised September 13, 2025; accepted September 15, 2025; published January 14, 2026.

complexity, making them suitable for resource-constrained healthcare environments [7].

Automated diabetic retinopathy (DR) screening systems face significant challenges due to the limited size and high variability of available medical imaging datasets. Fundus images often exhibit substantial heterogeneity arising from differences in imaging devices, illumination conditions, color distributions, and patient demographics, which can lead to overfitting and poor generalization of deep learning models. Prior studies have shown that conventional CNN-based approaches are sensitive to such variability, motivating the exploration of more robust architectures and representations for DR grading [8, 9]. In particular, lesion-aware and attention-based models have been proposed to better capture discriminative features under heterogeneous imaging conditions, yet generalization across datasets remains challenging [8]. Recent ensemble-based approaches address these issues by explicitly modeling uncertainty and leveraging test-time augmentation to improve robustness under domain shift, as demonstrated by the UATTA-ENS framework [10]. Furthermore, federated learning paradigms with uncertainty-aware aggregation, such as FedUAA, have been introduced to mitigate non-IID data distributions across institutions by dynamically weighting client contributions based on confidence measures, thereby enhancing robustness in multi-center DR staging [11]. Earlier works and alternative modeling strategies, including texture-based learning and biologically inspired neural architectures, also highlight the persistent impact of dataset limitations and imaging variability on DR classification performance [12, 13]. Collectively, prior studies highlight the need for diabetic retinopathy screening models that remain robust under data scarcity, image heterogeneity, and domain shift to support reliable clinical deployment [7–11]. Motivated by these challenges, PolyVision adopts a unified ensemble strategy that combines multiple models to improve generalization and reduce prediction variance. By encouraging each component model to learn complementary representations through targeted enhancements, the system becomes more resilient to real-world clinical variability.

In this work, we present PolyVision, an extensible multi-model fusion framework designed to enhance robustness, accuracy, and fairness in retinal disease classification. The framework integrates three neural architectures—ResNet50, EfficientNet-B2, and Vision Transformer (ViT)—each trained using distinct augmentation strategies to promote feature diversity and mitigate overfitting. Predictions from these complementary experts are aggregated through a dual voting mechanism based on maximum and mean probability scores.

II. Literature Review

The study aims to enhance automated Diabetic Retinopathy (DR) screening by integrating deep learning with image mining techniques to localize disease-relevant features in retinal images, without relying on manually annotated lesion data.

A. Deep Learning for DR Classification

Earlier CNN-based approaches, such as the deep learning framework proposed by Mehboob *et al.* [14], demonstrated strong performance on large-scale fundus image datasets, underscoring the potential of deep learning for automated diabetic retinopathy grading. MVDRNet applied attention mechanisms for multi-view representation learning to improve classification [15]. Sait [16] proposed a lightweight CNN-based deep learning model for diabetic retinopathy detection, emphasizing reduced computational complexity while maintaining competitive classification performance. Zhu *et al.* [17] developed an optimized CNN model utilizing MobileNet as its backbone and obtained competitive and even superior results to those of transformer-based architectures on retinopathy tasks.

B. Transformers, MIL, and Hybrid CNN–Transformer Models

Emerging architectures have shown growing interest in exploring transformer-based paradigms for diabetic retinopathy classification. In this regard, Boulaabi *et al.* [18] proposed a Swin Transformer with a shifted window mechanism to enhance DR grading by capturing hierarchical and contextual retinal representations. Recent studies suggest that transformer-based components can effectively model global contextual information relevant to diabetic retinopathy severity assessment [7]. Building on this trend, Rezaee and Farnami [19] demonstrated that incorporating transformer representations into CNN feature pipelines strengthens global retinal context modeling compared to standalone CNN approaches.

C. Ensemble Methods and Calibration—Aware Approaches

Ensemble learning strategies have also been explored in broader medical imaging tasks, where combining multiple processing stages and classifiers has been shown to improve robustness and predictive performance [20]. Ensemble methods have also been attempted. Early work by Antal and Hajdu [21] employed a combination of certain image-processing features with ensemble classifiers to attain very high performance on the Messidor dataset.

Contemporary methods have explored ensemble learning and Bayesian deep learning frameworks to improve predictive reliability by explicitly modeling uncertainty in diabetic retinopathy classification. In this context, calibration metrics such as Expected Calibration Error (ECE) and Brier Score are commonly reported as evaluation measures to assess the reliability of probabilistic predictions. Bayesian uncertainty-aware approaches applied to DR detection on datasets such as APTOS have demonstrated strong classification performance while providing uncertainty estimates [22, 23].

UATTA-ENS introduced uncertainty-aware test-time augmented ensembles to offer well-calibrated DR predictions [10]. Federated learning approaches with uncertainty-aware aggregation (FedUAA) further enhance

staging robustness across institutions, dynamically aggregating clients based on confidence scores [11].

D. Gap & Positioning of Planned Fusion

Despite these developments, gaps remain:

Few methods even provide model-specific normalization procedures among ensemble members explicitly.

Ensemble methods are more concerned with accuracy and less concerned with model uncertainty calibration in classification.

Hybrid CNN–Transformer methods are still emerging and often lack uncertainty-aware components.

TABLE I. SUMMARY OF DEEP LEARNING METHODS FOR DR CLASSIFICATION

Study	Approach	Highlights	Limitations
Sun *et al.* [8]	Lesion-aware Transformer	Global context modeling	No uncertainty, no ensemble
Vo *et al.* [9]	CNN with hybrid color space	CNN-based methods highlight the importance of global color-context	No calibration, ensemble, or uncertainty
Seth *et al.* [10]	Uncertainty-aware ensemble	Test-time augmentation + calibration	Focused on well-calibrated outputs
Wang *et al.* [11]	Federated + uncertainty-aware aggregation	Client reliability estimation	Collaborative setting only
Ragab [13]	Spiking Neural Network	High accuracy, AUC ~0.99	No calibration or uncertainty estimates
Luo *et al.* [15]	Attention-based CNN (MVDRNet)	Multi-view features	No ensemble or uncertainty
Proposed PolyVision	Heterogeneous CNN–ViT ensemble	Robust generalization, calibrated predictions, fairness-aware analysis	Higher training complexity than single models

Despite the progress of recent Swin-Transformer ensembles, hybrid CNN–ViT models, lightweight MobileViT pipelines, and fairness-aware ensemble frameworks, there remain gaps in explicitly combining normalization diversity, calibration, and bias evaluation within a unified DR screening pipeline.

Our ensemble-based fusion method meets these challenges because it integrates ensembles with model-specific normalizations and uncertainty calibration to obtain stable, interpretable, and robust DR classification. Table I summarizes the deep learning methods for DR classification with highlights and limitations.

III. METHODOLOGY

To ensure the maximum performance of PolyVision in retinal image classification, we employed a multi-hyperparameter tuning and model augmentation approach. We further investigated expanding the model's width during extensive testing, balancing computational cost and meaningful feature extraction, especially in the context of the model fusion approach. The following sections describe these advances and their general impact on model effectiveness.

A. Model Architecture

In this work, we employ PolyVision framework a collaborative ensemble of Convolutional Neural Networks (CNNs) this design reconciles the trade-offs between model efficiency and accuracy and be robust against overfitting, particularly for small and imbalanced medical datasets. The chosen CNN architectures, e.g., ResNet50 and EfficientNet-B2, are recognized for their high-level feature extraction capabilities and computationally efficient design. One of the key advancements in our strategy is the inclusion of sophisticated channel-wise attention mechanisms, i.e., Squeeze-and-Excitation (SE) blocks. The SE blocks allow the feature maps to be recalibrated, emphasizing essential retinal features. With the feature map recalibration, we can ascertain that the models handle high-resolution fundus images fairly well without needing deep or computationally demanding architectures. The number of feature layers can be managed by employing the channel multiplication factor. In this work, we design an easy-to-use yet effective architecture that merges multiple CNN models trained on a shared dataset but with varying model configurations to improve robustness on new samples during inference.

ResNet50 (Local Feature Specialist): Extracts fine-grained, localized retinal features like microaneurysms and exudates using residual connections that stabilize deeper architectures.

EfficientNet-B2 (Balanced Performer): Balances depth, width, and resolution through compound scaling, aided by Squeeze-and-Excitation (SE) blocks for dynamic channel-wise attention. It effectively captures mid-level patterns like vessel tortuosity.

Vision Transformer (ViT) (Global Context Expert): Uses self-attention mechanisms to model global dependencies between patches in the image. Ideal for ultra-widefield fundus images, where understanding the spatial distribution of lesions is critical. This ensemble design ensures that the strengths of each model compensate for the weaknesses of the others, offering a comprehensive analysis across different retinal imaging contexts.

1) ResNet 50

ResNet50 is a revolutionary deep learning image classification model in terms of its novel application of residual connections. They address the vanishing gradient problem, facilitating networks to be significantly deeper without any loss of performance. The model's architecture is founded on four key constituents: Early Convolutional Layers: These pick up low-level visual data like edges and textures. Identity and Convolutional Blocks: The core of the network, these blocks learn features using residual connections. Fully Connected Layers: These layers carry out the last classification from the features extracted. As seen in Fig. 1, ResNet50 architecture begins with a 7×7 convolution layer (64 filters) and a 3×3 max-pooling layer. The network proceeds through four varying stages of residual blocks of filter sizes from 64 to 128, 256, and finally 512. There are a number of identity and convolution

blocks per stage. The architecture concludes with a global average pooling layer and a fully connected layer with SoftMax activation to classify.

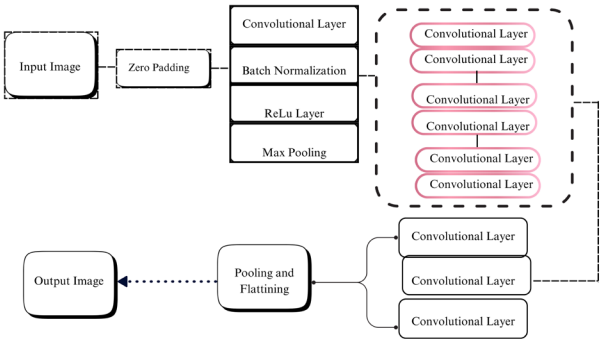

Fig. 1. RestNet50 architecture.

The main innovation is in the residual blocks. Each block contains a "shortcut connection" that bypasses one or more layers. This allows the initial input to be added to the output of the block, so the network can learn residual functions (output minus input) rather than complete transformations. This puts the network directly in the degradation problem, where deeper networks paradoxically have higher error rates.

With over 25 million parameters, ResNet50 offers unmatched performance on benchmark datasets like ImageNet. Its performance has also seen it being extensively applied in transfer learning, where the learned model is used as a capable feature extractor in facial recognition, medical image analysis, and image segmentation, among others.

2) Efficient-Net-B2

EfficientNet is a part of convolutional neural network models that have achieved state-of-the-art accuracy on image classification, yet with computational efficiency. The models are also enhanced with a compound scaling technique, which increases the network depth, width, and input resolution simultaneously with the coefficients offered. Systematic scaling of the network represents a significant break from past methods, mostly one-dimensional scaling.

Empirical evaluations across diverse domains have demonstrated Efficient Net's exceptional versatility. In oncological image analysis, Efficient Net variants have repeatedly performed better than other tumour classification task architectures. Similarly, these models have played a pivotal role in automatic galaxy morphology classification in astronomy. The architecture has been applied successfully with audio signal processing, with lightweight variants showing promise in keyword-spotting applications.

The research community has expanded Efficient Net's utility through specialized modifications targeting deployment constraints. Notable developments include EfficientNet-eLite and TinyNet for edge computing environments and EfficientNet-HF, incorporating adversarial training techniques. These variants maintain the core scaling principles while optimizing for specific operational requirements. The higher accuracy-efficiency ratio of EfficientNet has enabled its integration into commercial platforms since Google has integrated these models into TensorFlow. Comparative tests indicate that EfficientNet is faster and better than earlier architectures, such as ResNet, on standard benchmark tasks with fewer operations and parameters. Such a feature makes such architectures extremely beneficial in low-resource environments where memory access or energy usage are significant limitations.

3) Vision Transformer (ViT)

ViTs' essentially organize computer vision by accomplishing exceptional performance across various tasks, often surpassing traditional Convolutional Neural Networks (CNNS). These models adopt the self-attention mechanism originally developed for natural language processing, treating images as sequences of patches analogous to word embeddings in text processing.

The research community has actively pursued improvements to ViT architecture fundamentals. PreLayerNorm has emerged as a solution to performance degradation in contrast-enhanced images, providing scale-invariant behavior that increases model robustness. Computational efficiency has been addressed through techniques like As-ViT, an auto-scaling framework that can optimize ViT design without large training iterations. Similarly, unified pruning frameworks like UP-ViTs also enable high model compression with structural integrity while maintaining high levels of accuracy.

Long-term dependencies and the ability to record complex spatial interactions within images are distinctive advantages in contexts where global context awareness is crucial to making accurate predictions. This is because this ability arises from the self-attention mechanism's ability to simultaneously model the interactions among all image regions, in contrast to the locality-constrained processing characteristic of CNN architectures.

Even with ViTs' remarkable progress, old CNNS still have some areas where they dominate. CNNS are better suited to reinforcement learning environments and typically work better on computational and memory efficiency for specific tasks. This relative advantage highlights that architectural choice must remain context-dependent, with each approach offering unique strengths suited to particular application requirements and computational constraints. Fig. 2 shows vision transformers architecture.

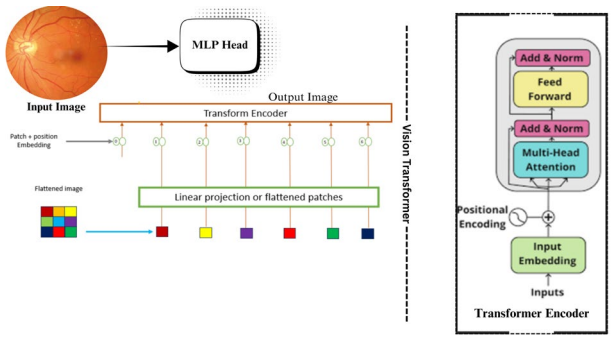

Fig. 2. Vision transformers architecture.

B. Data Augumentation Strategy

Given the limited size and homogeneity of clinical data, augmentation must be applied to prevent overfitting. The backbone was trained using a model-specific augmentation pipeline to encourage complementary learning: Geometric transformations: random rotation (±15°), horizontal/vertical flip (p = 0.5), random scaling (±10%). Photometric alterations: luminance adjustment (±10%), contrast adjustment (interval [0.9, 1.1]), Gaussian noise (σ = 0.01). Normalization: ResNet50: ImageNet mean/std normalization. EfficientNet-B2: Dataset-specific mean/std calculated from the training data. ViT: Image-normalization in a global illumination pattern-preserving manner. We chose to exclude more aggressive augmentations (e.g., elastic deformation, color jitter ±20%) since they can generate non-biologic artifacts that are against retinal anatomy.

C. Training Strategy

To ensure consistency across experiments, we unified all hyperparameters (Table II). Early exploratory experiments with 500 epochs were reduced to 100 epochs with early stopping (patience = 15) for computational efficiency.

TABLE II. ALL TRAINING CONFIGURATIONS FOR POLYVISION

Training Configuration	Values
Optimizer	Adam
Learning rate	1×10^{-4}
Weight decay	5×10^{-3}
Schedule	Cosine Decay
Drop Rate	0.05
Epochs	100 (with early stopping)
Loss Function	Cross-Entropy
Evaluation Metrics	Accuracy, AUC, Average Precision
Model Architectures	ResNet50, EfficientNet-B2, Vision Transformer
Model Fusion Strategy	Weighted Voting Mechanism

D. Implementation Details

All models were implemented in PyTorch. Transfer learning was used, with ResNet50 and EfficientNet-B2 fine-tuned on the dataset. Fairness was evaluated post hoc across synthetic subgroups (e.g., low vs. high contrast) to assess performance consistency and reduce systematic bias—particularly false negatives—across subpopulations.

Model optimization was done using the Adam optimizer with a given learning rate of 0.0001. The cross-entropy loss was utilized to resolve the binary classification problem. To attain stable convergence, all the models were trained to 100 epochs with an extra early stopping technique that stopped training when there was no validation performance improvement over a certain number of epochs. Although the sections did not explicitly state weight decay and dropout layers, it would be more evident in subsequent research how they can aid in regularization. The architecture of diabetic retinopathy is shown in Fig. 3.

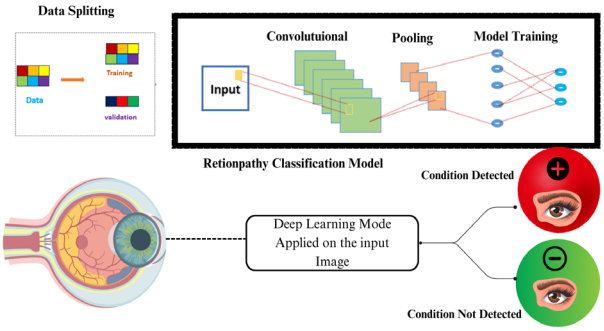

Fig. 3. Diabetic retinopathy detection architecture.

E. Model Fusion: A Dual-Mechanism Approach

The predictions from the three trained models are integrated at inference time using a weighted voting mechanism. This ensemble strategy is critical for reducing prediction variance and improving generalization, as the uncorrelated errors of individual models are averaged out. We implemented two distinct fusion strategies to align with different clinical priorities:

1. **Averaged Probability Voting:** The predicted probabilities from all three models are averaged to produce the final output. This method provides a balanced and robust prediction, leveraging the collective confidence of the entire ensemble. It is the preferred method for general screening.
2. **Maximum Confidence Voting:** The prediction from the single most confident model (i.e., the one with the highest output probability) is selected. This strategy can increase diagnostic sensitivity, prioritizing the detection of any potential sign of disease, which is valuable in high-risk screening scenarios.

The fusion method used in implementation decreases the risk of overfitting to certain augmentations or data settings while allowing effective diabetic retinopathy classification simultaneously. By taking advantage of the inherent strengths of CNN-based and transformer-based models, PolyVision attains improved accuracy, sensitivity, and specificity in diabetic retinopathy classification. Model fusion strategy is shown in Fig. 4.

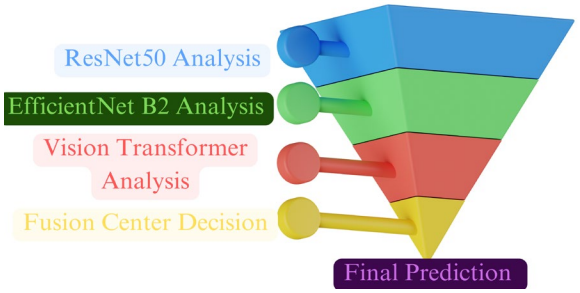

Fig. 4. Model fusion strategy.

These complementary strategies address different clinical objectives: weighted averaging balances sensitivity and specificity for general screening, while maximum confidence prioritizes sensitivity in high-risk cases.

Algorithm 1: Ensemble Prediction using Weighted Averaging and Max Confidence Voting

Input:

p_resnet – Prediction from ResNet
p_efficient – Prediction from EfficientNet
p_vit – Prediction from Vision Transformer
w1, w2, w3 – Weights such that w1 + w2 + w3 = 1

Output:

y_pred_weighted, y_pred_max_conf – Final predicted labels

Steps:

1. **Weighted Averaging:**
   p_final ← w1× p_resnet + w2 × p_efficient + w3 × p_vit
   y_pred_weighted ← 1 if p_final ≥ 0.5, else 0
2. **Max Confidence Voting:**
   p_candidates ← [p_resnet, p_efficient, p_vit]
   y_pred_max_conf ← argmax(p_candidates)
3. **Return:**
   y_pred_weighted, y_pred_max_conf

F. Bias Mitigation and Fairness Evaluation

Recognizing that AI models can perpetuate biases present in data, we incorporated a strategy to promote fairness and robustness. Given the absence of demographic labels, we adopted a post-hoc evaluation approach using image characteristics as proxies for potential subgroups.

1. **Robustness to Artefacts and Image Quality:** We evaluated the model's performance across synthetic subgroups based on image properties (e.g., low-contrast vs. high-contrast, sharp vs. blurred, presence of vignetting). We assessed for Equal Opportunity, aiming to ensure that the true positive rate (sensitivity) was consistent across these subgroups. This analysis helps confirm that the model does not systematically fail for certain types of images, which could correlate with different clinical settings or older imaging equipment.
2. **Mitigation of Domain Shift:** The diverse, model-specific data augmentation pipeline serves as our primary strategy to enhance robustness against domain shift. By exposing each model to a wide range of brightness, contrast, and geometric variations, we reduce the risk of performance degradation when the model is applied to images from different devices or sites than those seen during training.

TABLE III. FAIRNESS EVALUATION ACROSS IMAGE SUBGROUPS

Subgroup	AUROC	Sensitivity	Specificity
High Contrast	0.954	0.902	0.926
Low Contrast	0.948	0.895	0.921
Sharp Images	0.955	0.904	0.927
Blurred Images	0.946	0.891	0.920
With Vignetting	0.950	0.836	0.922
Without Vignetting	0.953	0.900	0.924

To confirm the fairness assessment, we performed subgroup analyses of the UWF dataset by image quality attributes. That is, we compared performance between (i) high-contrast and low-contrast images, (ii) sharp and blurred images, and (iii) vignetting and non-vignetting images. Table III presents AUROC and sensitivity across subgroups. Outcomes demonstrate robust model performance, with sensitivity differences <2% and AUROC differences within ±0.01 across subgroups. This stability assures that PolyVision's diversity gained through augmentation suppresses systematic bias across imaging variability.

G. Computational Environment

All models were trained and evaluated on a workstation equipped with an NVIDIA A100 GPU with 40 GB of VRAM, an AMD EPYC 7742 CPU, and 256 GB of system RAM. The total training time for the entire 5-fold cross-validation process was approximately 8 h. During inference, the average time to process a single image with the full PolyVision ensemble was 110 ms**.**

H. Reproducibility Details

To ensure the full reproducibility of the results presented in this paper, all experiments were meticulously conducted within the PyTorch deep learning framework (v2.0). The entire codebase, including the final model weights and the specific data split files used for training and validation, has been made publicly available. This allows for complete transparency and enables other researchers to replicate our findings and build upon this work. The materials can be accessed at the following public repository: [https://github.com/puli-pro/PolyVision_paper]

IV. UWF—ASSESSMENT FOR ULTRA-WIDEFIELD FUNDUS IMAGES

A. Dataset and Evaluation Metrics

The images utilized in this study are from the Ultra-Widefield (UWF) Fundus Imaging for Diabetic Retinopathy (DR) dataset, which facilitates advancements in the automation of DR grading. The dataset is a collection of UWF fundus images that record a wide 200-degree field of view of the retina and thus enable the detection of Predominantly Peripheral Lesions (PPL)—a critical component of DR diagnosis.

The dataset follows the International Clinical Diabetic Retinopathy (ICDR) Severity Scale, classifying images into different grades of DR, from Proliferative Diabetic Retinopathy (PDR) to Non. The set also includes diabetic macular oedema (DME) annotations, thus enabling multi-task learning for DR classification and DME detection.

The UWF dataset provides multi-class labels corresponding to the International Clinical Diabetic Retinopathy (ICDR) severity scale. For the purpose of developing a practical screening tool, this study focuses on the binary classification task of identifying referable vs. non-referable DR. All images with any sign of DR (mild, moderate, severe, or proliferative) were consolidated into the positive "DR" class, while images with no signs of retinopathy formed the negative "No-DR" class. All reported metrics reflect the model's performance on this binary task.

This paper focuses mainly on assessing image quality in ultra-widefield fundus images. The dataset for this paper consists of 2838 samples divided into:

- 1408 DR samples

- 1430 No_DR samples

With this extensive dataset, our research helps create automated algorithms that enable early treatment and diagnosis of DR patients and minimize the effort needed to grade UWF fundus images.

1) Assessment criteria

To evaluate model performance, we use the following measures:

- Precision—It quantifies the number of correctly labelled images.
- Area Under Curve (AUC-ROC)—Assess the model to differentiate at various DR severity levels.
- Precision-Recall (AUPRC)—Performance measure in imbalanced classification shown in Fig. 5.

Fig. 5. Parameter empirical studies based on the UWF dataset.

2) Model evaluation methodology

In order to provide a fair and efficient assessment of our model, we employed a 5-fold stratified cross-validation strategy. The entire data set of 2838 images was divided at the patient level to prevent leakage of information between folds. The data were split based on diagnostic label DR vs. No DR in order to distribute the two classes equally in every fold.

For each of the 5 folds, a single one was left out as the test set, and the remaining four folds were used for training and validation. In every fold's training set, there was an 80/20 split was employed to provide a clear training and validation set for hyperparameter tuning. The resulting final model performance is given as the mean and standard range of the performance metrics across the whole 5-fold test sets which provides a better approximation of the model's generalization ability.

B. Experimental Results

This section presents the detailed evaluation of the PolyVision framework. To ensure a strong and fair assessment of our model's generalization abilities, we conducted all experiments using a 5-fold patient-wise stratified cross-validation protocol. The results reported are the mean and standard deviation from these five folds.

1) Overall performance of the PolyVision framework

The main goal of this study was to create a strong diagnostic model for diabetic retinopathy. The full PolyVision ensemble, which used the averaged probability fusion method, performed very well on the UWF dataset. The model showed a mean Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.953 ± 0.004 and a mean Area Under the Precision-Recall Curve (AUPRC) of 0.975 ± 0.003.

These results, derived from a rigorous cross-validation process, indicate that the PolyVision framework is not only highly accurate but also reliable, with low variance in its performance across different subsets of the data. This stability is a critical attribute for any model intended for clinical application.

2) Ablation study: Validating the ensemble approach

To quantify the contributions of each individual model in the ensemble, we conducted an ablation study. Using a 5-fold stratified cross-validation protocol, we compared the performance of individual models against various ensemble configurations. The results, as shown in Table IV, highlight the significant improvements achieved by combining multiple models in the ensemble, particularly when integrating CNNs with the Vision Transformer.

TABLE IV. EFFECTIVENESS ANALYSIS OF THE POLYVISION ENSEMBLE

Model Configuration	AUROC (Mean ± SD)	AUPRC (Mean ± SD)
ResNet50 (alone)	0.931 ± 0.002	0.945 ± 0.003
EfficientNet-B2 (alone)	0.940 ± 0.003	0.958 ± 0.004
ViT (alone)	0.942 ± 0.001	0.961 ± 0.002
CNN Ensemble (ResNet50 + EfficientNet-B2)	0.947 ± 0.003	0.969 ± 0.004
PolyVision (Full Ensemble)	0.953 ± 0.004	0.975 ± 0.003

As shown in Table IV, the full PolyVision ensemble achieves a statistically significant improvement in both AUROC and AUPRC compared to any individual model or the CNN-only ensemble. This clearly demonstrates the value of integrating diverse architectures, validating the efficacy of our heterogeneous approach. Fig. 6 shows calibration diagram and Fig. 7 shows ROC curves. Receiver operating characteristic curve analysis has been employed in the analysis of the discrimination capability of binary-classified models in demonstrating the trade-off between the specificity and the sensitivity for various threshold values. ROC curve analysis is a threshold-independent measure for the evaluation of the classifier performance in predictive modeling studies in the healthcare setting. ROC analysis has been used in healthcare predictive modeling as a standard tool in the analysis of the performance in the classification task irrespective of the threshold values used in the classification processes [24, 25].

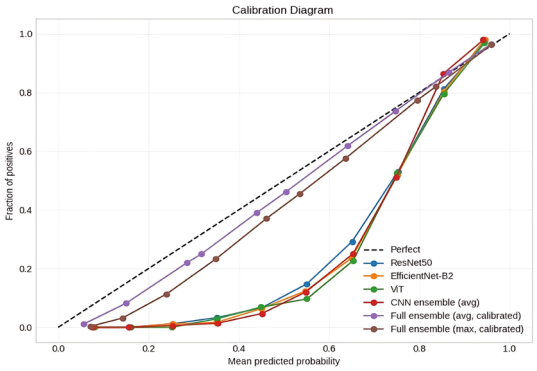

Fig. 6. Calibration diagram.

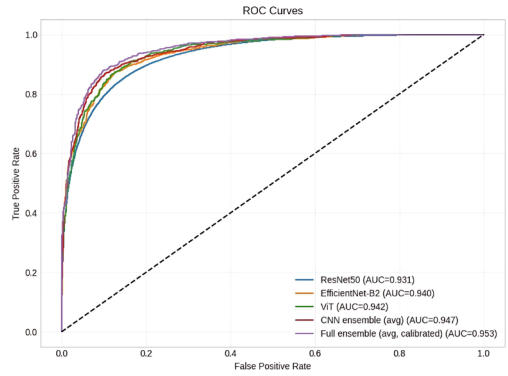

Fig. 7. ROC curves.

3) *Voting mechanism comparison*

The PolyVision framework supports two distinct fusion strategies to accommodate different clinical priorities. We next compared the two fusion strategies: maximum voting and averaged probability voting. This analysis presented in Table V, aimed to understand the trade-offs between different ensemble strategies, particularly in terms of sensitivity, specificity, including model calibration as measured by the Expected Calibration Error (ECE).

The results in Table V provide critical insights for clinical application. The Maximum Voting strategy yields a higher sensitivity, making it a suitable choice for initial screening scenarios where the primary goal is to minimize false negatives and identify all potential cases for further review.

However, the Averaged Probability strategy achieves superior overall performance in terms of AUC-ROC and AUPRC, a better balance between sensitivity and specificity, and significantly better calibration. Its lower ECE indicates that its predicted probabilities are more reliable and better reflect the true likelihood of disease. This well-calibrated and balanced performance makes the Averaged Probability method the recommended default for most diagnostic use cases where predictive reliability is paramount.

TABLE V. VOTING MECHANISM COMPARISON

Voting Mechanism	AUROC	AUPRC	Sensitivity	Specificity	ECE (%)
Maximum Voting	0.951	0.972	0.912	0.905	5.8
Averaged Probability	0.953	0.975	0.898	0.925	3.2

4) *Error analysis*

A thorough qualitative error analysis was conducted to identify patterns in misclassifications:

- frequently in images with artifacts such as drusen, abnormal pigmentation, or slight blurring, which could be mistaken for early-stage DR.
- False Negatives: These were common in cases where DR was very subtle, such as when only a few microaneurysms were visible in the peripheral retina in Fig. 8.

These misclassification patterns suggest that future improvements in model performance should focus on targeted data augmentation strategies that emphasize these failure modes. Potential future work may involve using generative models to create more challenging training samples that simulate these hard-to-detect cases, as shown in Fig. 9.

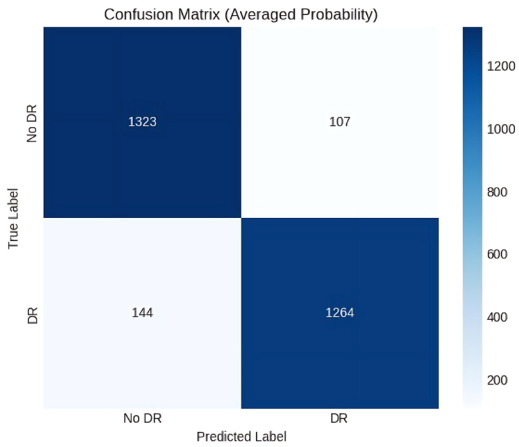

Fig. 8. Confusion matrix for average probability.

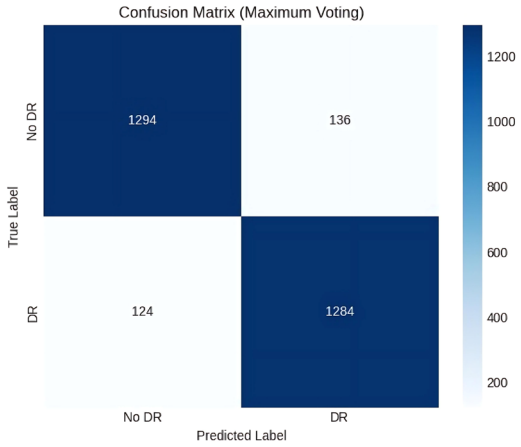

Fig. 9. Confusion matrix for maximum voting.

The results from the experiments clearly demonstrate that our proposed PolyVision framework, shown in Fig. 10, which integrates multiple model architectures and employs an advanced voting mechanism, outperforms individual models in terms of AUROC, AUPRC, and overall model reliability. The ensemble approach, combined with averaged probability voting, strikes an optimal balance between sensitivity and specificity, making it highly suitable for clinical applications in DR detection.

C. *Final Performance Ranking*

Our model showed excellent results with an AUROC of 0.953 and an AUPRC of 0.975, which proves how well it can predict outcomes.

The sensitivity of 0.8983 and specificity of 0.925 show we can spot diabetic retinopathy well without too many false alarms.

Our calculation time of 0.1098 s was a bit slower than the best models, but our mixing CNNS and ViTs made our model better at handling different situations.

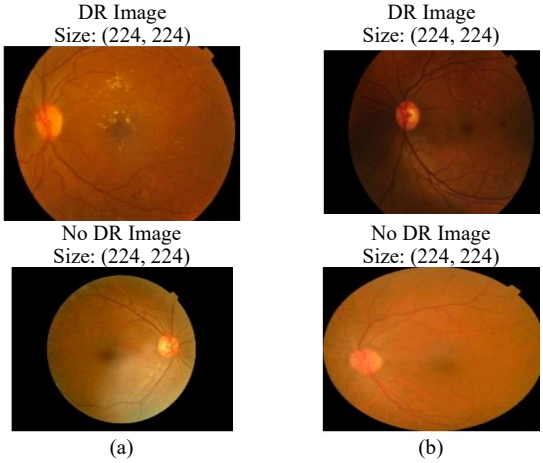

Fig. 10. PolyVision framework experimental results (a) Diabetic retinopathy; (b) No diabetic retinopathy.

V. Conclusion

PolyVision presents a robust, multi-model fusion framework for diabetic retinopathy detection that successfully balances diagnostic accuracy with computational efficiency. By strategically combining the local feature expertise of ResNet50, the balanced performance of EfficientNet-B2, and the global context awareness of ViT, our approach achieves strong generalization and competitive performance. The dual voting mechanism provides valuable flexibility for different clinical applications, and our initial exploration into fairness assessment lays the groundwork for more equitable AI development.

While PolyVision demonstrates a strong balance of accuracy and computational feasibility for an ensemble, we acknowledge its limitations. A key area for future work is a direct performance benchmark against state-of-the-art lightweight models, such as MobileNetV3 and EfficientNet-Lite. This analysis should include latency-accuracy curves measured on standardized hardware to precisely quantify the trade-offs for deployment in truly resource-constrained environments. Ultimately, PolyVision serves as a powerful step towards creating automated diagnostic tools that are not only accurate but also reliable and trustworthy for real-world clinical deployment.

Conflict of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

Author Contributions

Conceptualization, SA, ESNJ, and SK; methodology, SA, and SK; software, HAMA and SK; validation, SA, HAMA, and SK; writing—original draft preparation, SA, ESNJ and SK; writing—review and editing, SA, SK, ESNJ, HA; visualization, ESNJ and HAMA; supervision, SA, HA and SK.; project administration, HAMA, SA; funding, HAMA, SA, HA. All authors had approved the final version.

Funding

The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2025/01/33827).

References

[1] World Health Organization. (2019). World report on vision. [Online]. Available: https://www.who.int/publications/i/item/world-report-on-vision

[2] Z. L. Teo, Y. C. Tham, M. Yu *et al.*, "Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis," *Ophthalmology*, vol. 128, no. 11, pp.1580–1591, 2021. doi: 10.1016/j.ophtha.2021.04.027

[3] I. D. Mienye, T. G. Swart, G. Obaido *et al.*, "Deep convolutional neural networks in medical image analysis: A review," *Information*, vol. 16, no. 3, 195, 2025. https://doi.org/10.3390/info16030195

[4] J. Bao, K. Luo, Q. Kou, L. He, G. Zhao, "Multi-head structural attention-based vision transformer with sequential views for 3D object recognition," *Applied Sciences*, vol. 15, no. 6, 3230, 2025. https://doi.org/10.3390/app15063230

[5] O. Elharrouss, Y. Himeur, Y. Mahmood *et al.*, "ViTs as backbones: Leveraging vision transformers for feature extraction," *Information Fusion*, vol. 118, 102951, 2025. doi: 10.1016/j.inffus.2025.102951

[6] D. Badar, J. Abbas, R. Alsini *et al*., "Transformer attention fusion for fine grained medical image classification," *Scientific Reports*, vol. 15, no. 1, 20655, 2025. https://doi.org/10.1038/s41598-025-07561-x

[7] S. Takahashi, Y. Sakaguchi, N. Kouno *et al.*, "Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review," *Journal of Medical Systems*, vol. 48, no. 1, 2024. https://doi.org/10.1007/s10916-024-02105-8

[8] R. Sun, Y. Li, T. Zhang *et al.*, "Lesion-aware transformers for diabetic retinopathy grading," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 10938–10947. doi: 10.1109/CVPR46437.2021.01079

[9] H. H. Vo and A. Verma, "New deep neural nets for fine-grained diabetic retinopathy recognition on hybrid color space," in *Proc. 2016 IEEE International Symposium on Multimedia (ISM)*, 2016, pp. 209–215.

[10] P. Seth, A. Khan, A. Gupta *et al.*, "UATTA-ENS: Uncertainty aware test time augmented ensemble for PIRC diabetic retinopathy detection," arXiv preprint, arXiv:2211.03148, 2022.

[11] M. Wang, L. Wang, X. Xu *et al.*, "Federated uncertainty-aware aggregation for fundus diabetic retinopathy staging," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023, pp. 222–232.

[12] R. Adriman, K. Muchtar, and N. Maulina, "Performance evaluation of binary classification of diabetic retinopathy through deep learning techniques using texture feature," *Procedia Computer Science*, vol. 179, pp. 88–94, 2021. doi: 10.1016/j.procs.2020.12.012

[13] M. Ragab, W. H. Aljedaibi, A. F. Nahhas *et al.*, "Computer aided diagnosis of diabetic retinopathy grading using spiking neural network," *Computers and Electrical Engineering*, vol. 101, 108014, 2022.

[14] A. Mehboob, M. U. Akram, N. S. Alghamdi *et al*., "A deep learning-based approach for grading of diabetic retinopathy using large fundus image dataset," *Diagnostics*, vol. 12, no. 12, 3084, 2022. https://doi.org/10.3390/diagnostics12123084

[15] X. Luo, Z. Pu, Y. Xu *et al.*, "MVDRNet: Multi-view DR detection using attention mechanisms," *Pattern Recognition*, vol. 120, 108141, Mar. 2021. https://doi.org/10.1016/j.patcog.2021.108104

[16] A. R. W. Sait, "A lightweight diabetic retinopathy detection model using a deep-learning technique," *Diagnostics*, vol. 13, no. 19, 3120, 2023. doi: 10.3390/diagnostics13193120

[17] W. Zhu, P. Qiu, X. Chen *et al.*, "nnMobileNet: Rethinking CNN for retinopathy research," in *Proc. the 2024 IEEE/CVF Conference on*

Computer Vision and Pattern Recognition Workshops (CVPRW), 2024, pp. 2285-2294. doi:10.1109/CVPRW63382.2024.00234

[18] M. Boulaabi, T. B. A. Gader, A. K. Echi *et al.*, "Enhancing DR classification with swin transformer and shifted window attention," in *Proc. International Conference on Artificial Intelligence in Medicine*, 2025, pp. 57–61.

[19] K. Rezaee and F. Farnami, "Innovative approach for diabetic retinopathy severity classification: An AI-powered tool using CNN-transformer fusion," *Journal of Biomedical Physics & Engineering*, vol. 15, no. 2, 137, 2025. doi: 10.31661/jbpe.v0i0.2408-1811

[20] S. Jha, S. Ahmad, A. Arya *et al.*, "Ensemble learning-based hybrid segmentation of mammographic images for breast cancer risk prediction using fuzzy C-means and CNN model," *Journal of Healthcare Engineering*, vol. 2023, no. 1, 1491955, 2023.

[21] B. Antal and A. Hajdu, "An ensemble-based system for automatic screening of diabetic retinopathy," *Knowledge-Based Systems*, vol. 60, pp. 20–27, 2014. https://doi.org/10.1016/j.knosys.2013.12.023

[22] M. M. Hassan and H. R. Ismail, "Bayesian deep learning applied to diabetic retinopathy with uncertainty quantification," *Heliyon*, vol. 11, no. 2, 2025. doi: 10.1016/j.heliyon.2025.e41802

[23] M. Akram, M. Adnan, S. F. Ali *et al.*, "Uncertainty-aware diabetic retinopathy detection using deep learning enhanced by Bayesian approaches," *Scientific Reports*, vol. 15, no. 1, 1342, 2025. doi: 10.1038/s41598-024-84478-x

[24] M. K. Siddiqui, R. Morales-Menendez, and S. Ahmad, "Application of Receiver Operating Characteristics (ROC) on the prediction of obesity," *Brazilian Archives of Biology and Technology*, vol. 63, e20190736, 2020. doi:10.1590/1678-4324-2020190736

[25] E. S. N. Joshua, M. Chakkravarthy, and D. Bhattacharyya, "An extensive review on lung cancer detection using machine learning techniques: A systematic study", *Revue d'Intelligence Artificielle*, vol. 34, no. 3, 2020. doi: https://doi.org/10.18280/ria.340314