

ARMDiaRD: A robust multi-class diabetic retinopathy detection using hybrid swin transformers with hierarchical fusion

J. Dhiviya Rose* and Ved Prakash Bhardwaj

School of Computer Science,
UPES, Bidholi, Dehradun 248007,
Uttarakhand, India

*E-mail: dhiviyarj@ddn.upes.ac.in

Received for publication
August 22, 2025

Abstract

Diabetic retinopathy (DR) is an eye issue of prolonged diabetes, which continues to be a major cause of vision loss among middle-aged adults. Early detection of DR through advanced imaging and artificial intelligence (AI) techniques can lead to a significant reduction in risk and severity of visual loss. This proposed ARMDiaRD combines state-of-the-art techniques from EfficientNet, Swin Transformers, and multi-scale feature fusion (MSFF) to enhance multiclass classification of DR severity levels across diverse generalised datasets. In the proposed architecture, a global context (GC) block is integrated into EfficientNet to capture long-range dependencies and contextual relationships. This is followed by Swin Transformer layers equipped with an MSFF block that hierarchically aggregates features from multiple levels, enabling the model to learn richer and more discriminative representations. The fundus images of four publicly available datasets for DR: asia pacific tele ophthalmology society (APTOS), Indian diabetic retinopathy image dataset (IDRID), MESSIDOR-V1, and EYEPACS, are combined through a comprehensive aggregation process to facilitate the training and testing of the generalised model. This design demonstrates consistent improvements across multiple evaluation metrics, underlining its potential to reduce misclassification in medical diagnosis. To evaluate the proposed model, the performance metrics, such as accuracy, precision, recall, specificity, and F1-score, are calculated along with quadratic weighted kappa (QWK), Spearman rank correlation coefficient, and mean absolute error (MAE). Simulation results revealed that the proposed model achieved an accuracy of 87.59%, precision of 87.6%, recall of 87.9%, QWK of 91.47% and Spearman coefficients with 92.53% respectively. Importantly, the MAE, a critical metric for evaluating false predictions in medical diagnosis, is 0.1736 in the proposed model. The outcome clearly demonstrates the supremacy of the proposed model over the other models in handling multiple larger datasets. This research shows the ubiquitous nature of the proposed model to predict the severity grades using larger complex datasets. This proposed ARMDiaRD confirms that combining local, global, and hierarchical characteristics is highly effective in preventing overfitting issues with the existing architecture, such as CNN, when tested using new data for reliable clinical image analysis.

Keywords

Diabetic retinopathy, fundus image, multi-class classification, Swin transformers, multi-scale feature fusion

I. Introduction

Diabetic retinopathy (DR) is a vision-specific problem triggered by long-term complications associated with diabetes mellitus (DM). It is considered to be an important cause of vision loss and eye impairment among middle-aged working adults. Long-term diabetes, if left untreated, can significantly impair or completely degrade an individual's vision by damaging blood vessels in the retina [1–3]. Thus, retinal blood vessels are highly susceptible to damage due to prolonged high blood sugar levels, which makes the blood vessels more likely to sustain further damage or even rupture [4]. The various risks of DR are influenced by blood sugar control, duration of diabetes, hypertension, genetic predisposition, and lipid abnormalities [5]. The most important thing to reduce the severity of DR, which is one of the irreversible blindness-causing scenarios in people, will be early detection and counseling.

Fundus photography refers to the process of capturing the various parts of the eyeball, the macula, the posterior pole, the retina, and the optic disk. The macula is the central point for detailed vision, and the posterior pole includes the macula region and the optic disk at the back of the eye. The retina is the light-sensitive layer, and the optic disk is the site where the retinal nerve fibers converge and exit the eye, creating a blind spot. These images, known as fundus images, are critical for detecting and monitoring a range of retinal conditions, including DR [6]. Fundus images are taken with the help of special cameras called fundus cameras, which combine a low-power microscope with a specialized flash-enabled camera. In fundus images, a retinal lesion is any abnormal change or damage observed in the tissues of the retina, typically visible in fundus photography [7, 8]. The development DR is closely associated with the appearance and accumulation of retinal lesions, most notably microaneurysms, hemorrhages, and exudates.

- *Microaneurysms* is the early stages of DR, where small and round red dots in the retina are caused by localized dilation of capillary walls.
- *Hemorrhages* are retinal bleeds that occur when weakened blood vessels rupture, appearing as dark red spots or patches within the retinal layers.
- *Exudates* are yellow-white lipid or protein deposits that leak from damaged retinal vessels, often indicating advanced retinopathy.

The International Clinical Diabetic Retinopathy (ICDR) scale is a globally accepted framework for DR

grading with retinal findings observed through fundus examination. The scale categorizes DR into five stages as stated below:

- *No DR/Class 0*, which is termed as having no complications.
- *Mild non-proliferative diabetic retinopathy (NPDR)/Class 1* defined by the presence of microaneurysms only.
- *Moderate NPDR/Class 2* is defined by the appearance of intraretinal hemorrhages, venous beading, and hard exudates.
- *Severe NPDR/Class 3* identified by the “4-2-1 rule,” which indicates extensive intraretinal hemorrhages in all four quadrants, venous beading in at least two quadrants, or prominent intraretinal microvascular abnormalities in at least one quadrant.

Proliferative diabetic retinopathy (PDR)/Class 4, which is the most advanced stage, is marked by neovascularization and/or vitreous hemorrhage, posing a significant risk of vision loss in the absence of timely progression of treatment.

The ICDR scale is widely adopted in both public health screening programs and machine learning-based diagnostic systems [9]. Artificial intelligence (AI) in healthcare has shown notable progress in classifying DR with binary classes (yes or no DR) and multiclassification (ICDR classifying to five classes), making the task automated, which greatly benefits public mass screening [10–13]. These systems aim to optimize early detection and improve patient care. The use of AI in predicting diabetes and related diseases has been a state-of-the-art research area that has been pursued for over a decade. The machine learning algorithms were predominantly used for predicting diabetic mellitus, considering various lifestyle parameters [14, 15]. For example, Bhat et al. [16] developed a machine-learning risk prediction pipeline combining multiple classifiers and imbalance-handling strategies, as in most of the diabetic datasets, class imbalance seems to be a major issue. Even though much state-of-the-art research has been proven on the publicly available datasets as given in Table 1, the use of local datasets for validating the experiments remains an open call for further researchers. The work done by Bhat et al. [17, 18] in developing a machine-learning prediction model stands as an example in working on the real data of the North Kashmir cohort. AI systems can efficiently analyze retinal images too, detect early stages of DR, and categorize severity levels, helping to diagnose and prioritize treatment. Advances in portable and smartphone-based fundus cameras

Table 1: Summary of major fundus image datasets relevant to the ICDR grading system

Dataset	Year	Images	Grade	Resolution	Country	Strengths	Limitations
EyePACS [19]	2015	88,702	5	Varying dimensions	USA	Large-scale dataset	Class imbalance
IDRiD [21]	2018	516	5	4,288 × 2,848	India	Classification and segmentation	Small size, limited masks
APTOS [20]	2019	3,662		512 × 512	Asia-Pacific	Clean, uniform images	Small dataset, class imbalance
FGADR [22]	2020	1,842	5	2,048 × 3,072	China	Balanced dataset	Moderate size, population-specific
MESSIDOR [23]	2008	1,200	4	Varying dimensions	France	Good for bench-marking	Uses only 4 grading
EOphtha [24]	2014	611	Not graded	2,544 × 1,696	France	Early DR detection	No severity grading
DDR	2020	13,673	5	High-resolution	China	Large-scale, diverse images	Variable image quality

DDR, diabetic retinopathy dataset; ICDR, international clinical diabetic retinopathy.

have also enabled widespread public screening, particularly in remote and underdeveloped regions, as these images provide a low-cost way of detecting early pathological changes, such as microaneurysms, hemorrhages, and exudates.

a. Problem statement

Deep learning architectures that use fundus images in DR detection have shown good results in automating the identification of various stages of DR, but challenges persist. In particular, the efficient handling of large datasets and the accurate classification of severity grades remain areas that require further research. As models become deeper, the risk of missing important characteristic correlations increases, and the model becomes overfitted, potentially degrading classification performance when applied to a new set of data [12]. Thus, extracting highly sensitive deep features from extensive datasets for severity classification remains a critical challenge.

b. Motivation and contribution of the research

Motivated by the identified challenges, this research proposes ARMDiaRD: a novel hybrid Swin Transformer model combined with improved marine predator optimization (IMPO) for multiclassification of

DR severity grades with a massive integrated global fundus dataset. The study also introduces global context (GC) capture with local feature concatenation using multi-scale feature fusion (MSFF), which enables deeper feature extraction and improves classification outcomes.

The primary contributions of this study are listed below.

1. Introduction of the hybrid model, which first enriches the convolutional neural network (CNN) features with GC attention through EfficientNet-B0 and feeds them into Swin layers that deploy MSFF.
2. Development of multiple Swin block architectures with MSFF for depth feature extraction, thereby enhancing classification performance by concatenating multiple features extracted from large datasets.
3. Use of a hyperparameter tuning algorithm, improved marine predator algorithm (IMPA), for better performance through searching for the optimized hyper-parameter space.
4. Extensive experimentation and result discussion carried out using the existing multiclassification fundus benchmark datasets, EyePACS, APTOS, MESSIDOR-V1, MESSIDOR-V2, and IDRiD. The proposed model has outperformed state-of-the-art models in comparison.

This paper is structured as follows. Section II provides an overview of the existing methods for the detection of DR and studies of various fundus data sets. Section III describes the proposed system methodology. Section IV presents the experimental setup and results, and Section V concludes the paper with future research hopes in the area of DR.

II. Related Works

In recent studies, diverse ophthalmic imaging techniques have been used to acquire an in-depth understanding of the severity of ocular irregularities by predicting the stage of DR (DR) with greater precision. This section details the numerous annotated fundus image datasets that have been developed to facilitate automated DR detection using deep learning approaches. The latter subsection is designed to give a rich state-of-the-art literature on the various existing models applied to the datasets. The literature is reviewed across benchmark datasets, such as APTOS, EyePACS, IDRiD, and MESSIDOR, which have taken place across the deep learning evolution from CNNs, hybrid models, vision transformers, and Swin transformer models.

a. Fundus datasets

Fundus photography is the cornerstone of the diagnosis and classification of DR, enabling non-invasive visualization of the retina to identify pathological features such as microaneurysms, hemorrhages, and exudates. Numerous annotated fundus image datasets have been developed to facilitate automated DR detection using deep learning approaches, as given in Table 1 with their image dimensions and nativity. The EyePACS dataset, popularized through a 2015 Kaggle competition, continues to be one of the largest publicly available DR datasets, comprising over 88,000 images, with 35,126 labeled fundus images annotated with five classes according to the ICDR DR grades [19]. APTOS 2019 dataset includes 3,662 preprocessed images with the same grading scale and was designed for high consistency in classification tasks [20]. The IDRiD dataset [21] is unique in offering both DR severity labels and fine-grained pixel-level segmentation of lesions, including microaneurysms, hemorrhages, hard exudates, and soft exudates. FGADR dataset [22] provides fine-grained annotations and segmentation masks over 1,842 high-resolution images, supporting advanced lesion-aware DR grading research. MESSIDOR, released in 2008, includes 1,200 images graded for DR severity and image quality [23]. Although it uses a 4-class grading

scheme, it is often mapped to the 5-class ICDR system for consistency. The diabetic retinopathy dataset (DDR) from China has 13,000 images with both DR and diabetic macular edema (DME) labels. The E-Ophtha dataset [24] provides images with pixel-level annotations for microaneurysms and exudates for lesion-focused research, and it lacks full DR severity grades. Together, these datasets form the foundation of modern DR classification and segmentation research, offering diverse combinations of image resolutions, annotation types, and patient demographics. They are frequently used to train and validate CNNs, Vision Transformers, and hybrid deep learning architectures.

b. Existing deep learning architectures

The state-of-the-art literature is reviewed across benchmark datasets, such as APTOS, EyePACS, IDRiD, ODIR, and MESSIDOR, as shown in Table 2, done across conventional CNNs, hybrid models, vision transformers, and Swin transformer models. The CNNs have played a pivotal role in advancing computer vision by solving various visual recognition challenges. A CNN is structured with multiple convolutional layers for feature extraction, each designed to learn and recognize specific features or patterns in the input image. These layers compute feature maps by sliding small filters across the image. Pre-trained CNN models may be further fine-tuned for different applications, leveraging knowledge gained from exposure to large, diverse datasets. Transfer learning is a process of fine-tuning pre-trained CNNs, such as VGGNet, ResNet, Inception, and DenseNet, on the targeted fundus dataset for further learning and feature selection. Early approaches relied heavily on CNN architectures, such as Inception-v3 and ResNet [25]. Gulshan et al. [19] pioneered a high-accuracy binary classification system for DR detection using Inception-v4 on Eye-PACS, reporting an AUC of 0.99. Porwal et al. [21] established the first benchmark using Inception-v3 on the IDRiD dataset, achieving around 78% accuracy. Islam et al. (2020) [26] advanced this by introducing a multitask CNN that jointly performed lesion segmentation and classification, attaining a Quadratic Weighted Kappa (QWK) score of 0.80. Pratt et al. (2016) [27] employed a 10-layer CNN to perform 5-class classification on EyePACS with approximately 75% accuracy. Other enhancements, such as the use of DenseNet-121 with focal loss (Wang et al., 2021) [28] and ResNet50 with domain adaptation (Voets et al., 2019) [29], tackled data imbalance and cross-domain generalization issues. Another research conducted by Tiwari et al. [30] also claimed that the MobileNet

Table 2: Literature review of various deep learning techniques for DR detection

Year	Model used	Dataset	Accuracy (%)	Key contributions	Study finding
2023 [31]	Resnet50	APTOS	83.90	Compared the model with VGG16, Xception, AlexNet and found Resnet outperformed	Performance can be improved
2021 [32]	Multi-scale attention network (MSA-Net)	EYEPACS and APTOS	84.40	MSA-Net helps to retrieve features at multiple scales	Performance can be improved
2022 [33]	Inception-V3	89,947 images from various datasets	99	Robust binary classification for fundus image quality	ICDR grading can be implemented
2023 [34]	Resnet50	APTOS	83.90	Compared the model with VGG16, Xception, AlexNet and found Resnet outperformed	Performance can be improved
2023 [35]	Inception-V3	APTOS	98.7	Image enhancement using CLAHE	Better performance with preprocessing
2023 [36]	DenseNet121	APTOS	97.30	Combines VGG16, XGBoost, and DenseNet121; highlights overfitting risk	Overfitting due to class 0 dominance
2023 [37]	SqueezeNet, Darknet-53, EfficientNet-B0	ODIR	95, 99.4, 90	Multi-classification: normal, glaucoma, cataract	ICDR DR grading can be extended
2024 [38]	DeepDR Plus (ResNet-50 + self-attention)	83,500 images from various datasets	~84.6	Predicts DR progression time; supports personalized screening	Time prediction solved; classification remains open
2025 [39]	Inception ResNet V2, MobileNet, Residual Net	645 clinical images	93	Binary classification of eye diseases	DR grading can be implemented
2025 [40]	Inception-ResNet-v2 + GRU	APTOS	98.00	FFO fine-tunes GRU	Optimization improves accuracy

CLAHE, Contrast Limited Adaptive Histogram Equalization; DR, diabetic retinopathy; FFO, Fennec Fox Optimization; GRU, gated recurrent unit; ICDR, international clinical diabetic retinopathy; MSA, multi-head self attention.

outperformed all the other architectures while testing in the fundus photography dataset [30] when tested with EyePACS. Dai et al. [38, 19] introduced DeepDR Plus, a deep learning system developed to estimate the time to progression of DR, a major cause of preventable vision loss. The model was pretrained on an extensive dataset comprising 717,308 fundus images from 179,327 individuals and validated using 118,868 images from a multiethnic cohort of 29,868

participants. DeepDR Plus utilized a pretrained ResNet-50 as a feature extractor, complemented by using self-attention to focus on diagnostically relevant areas in fundus images. It achieved strong predictive performance, with concordance indexes between 0.754 and 0.846. Although the model was trained primarily on a Chinese population, its generalizability could be enhanced through further training using more diverse clinical and demographic datasets.

c. Improved marine predators algorithm (IMPA)

Hyperparameter tuning using optimization algorithms helps in achieving good performance for deep learning models. Manual tuning or grid search is often inefficient, especially when the search space is large or when training is computationally expensive. To address this, optimization algorithms are employed to automate the hyperparameter search process more effectively. The improved marine predators algorithm (IMPA) is a population-centric optimization method, shares similarities with various other metaheuristic strategies, particularly with the enhanced marine predators algorithm (MPA) [41, 42]. A notable attribute of this algorithm is its periodic dispersion of the starting solution throughout the search domain, as articulated in Eq. (1).

$$X_0 = X_{\min} + \text{rand} \cdot (X_{\max} - X_{\min}) \quad (1)$$

where rand denotes a randomly generated vector with elements uniformly distributed between 0 and 1, while $X_{\min}$ and $X_{\max}$ represent the minimum and maximum bounds of the parameters being optimized, respectively. This initialization ensures that the initial population of candidate solutions is spread uniformly across the feasible search space, laying a solid foundation for the subsequent optimization phases. Each candidate in the population is represented as a vector of hyperparameter values, for example,

$$X_i = [\text{learning rate, batch size, depth, dropout, ...}].$$

For each candidate X_i , the model is trained using the corresponding hyperparameters (HP). After training, the model is evaluated on a validation set to obtain a performance metric. For classification tasks, commonly used metrics include validation accuracy or F1-score. This validation performance metric is then used as the fitness value for the candidate, depending on the problem X_i as mentioned in Eq. (2).

$$\text{fitness} = \begin{cases} \text{validation accuracy,} & \text{if maximization problem} \\ -\text{validation loss,} & \text{if minimization problem and fitness maximized} \\ \text{validation loss,} & \text{if minimization problem and fitness minimized} \end{cases} \quad (2)$$

IMPA finds the optimized hyperparameter solutions faster by using certain social dynamics and dynamic weights initialization, which can prevent ending up with local optima. IMPA is better at exploring global optima, improving overall performance. The update population X in IMPA update rules uses Levy flights and adaptive position adjustments inspired by marine predator behavior, balancing exploration and exploitation in the search space.

III. System Methodology

ARMDiaDR introduces a novel, generalized robust hybrid architecture framework for the detection of multiclass DR based on ICDR grading, which alternates Swin Transformer blocks with multiscale feature fusion layers and CNN layers, along with the baseline model EfficientNet-B0 with GC. The experimental system flow includes a robust data aggregation process, with hyperparameter selection and model creation, as shown in Figure 1 from multiple datasets, followed by the training and testing process.

a. Data agglomeration pipeline

ARMDiaDR system combines multiple data sources captured using different devices or imaging conditions. To gain a more comprehensive view, a strong data agglomeration process is carried out on the individual datasets before the model training process to enhance both image quality and model training efficiency. The entire pipeline consists of three main phases: preprocessing, augmentation, and quality-aware down-sampling. This process serves as an important feature, as images from various datasets (APTOS, IDRID, MESSIDOR-V1, MESSIDOR-V2, and EYEPACS) are merged for experimentation. The pipeline automatically handles the preprocessing of multiple datasets as shown in Figure 2, eliminating the need for manual intervention and ensuring consistency across datasets. The images are loaded from their respective datasets, which are preprocessed and aggregated, and the majority classes are subject to quality-aware down-sampling and augmented to generate the training and testing datasets.

The quality-based down-sampling strategy was incorporated to ensure that only diagnostically rich images were retained, particularly among healthy samples, where overrepresentation is common for generating the balanced dataset. Image quality was assessed using a set of objective metrics, including sharpness (edge clarity), contrast (luminance and

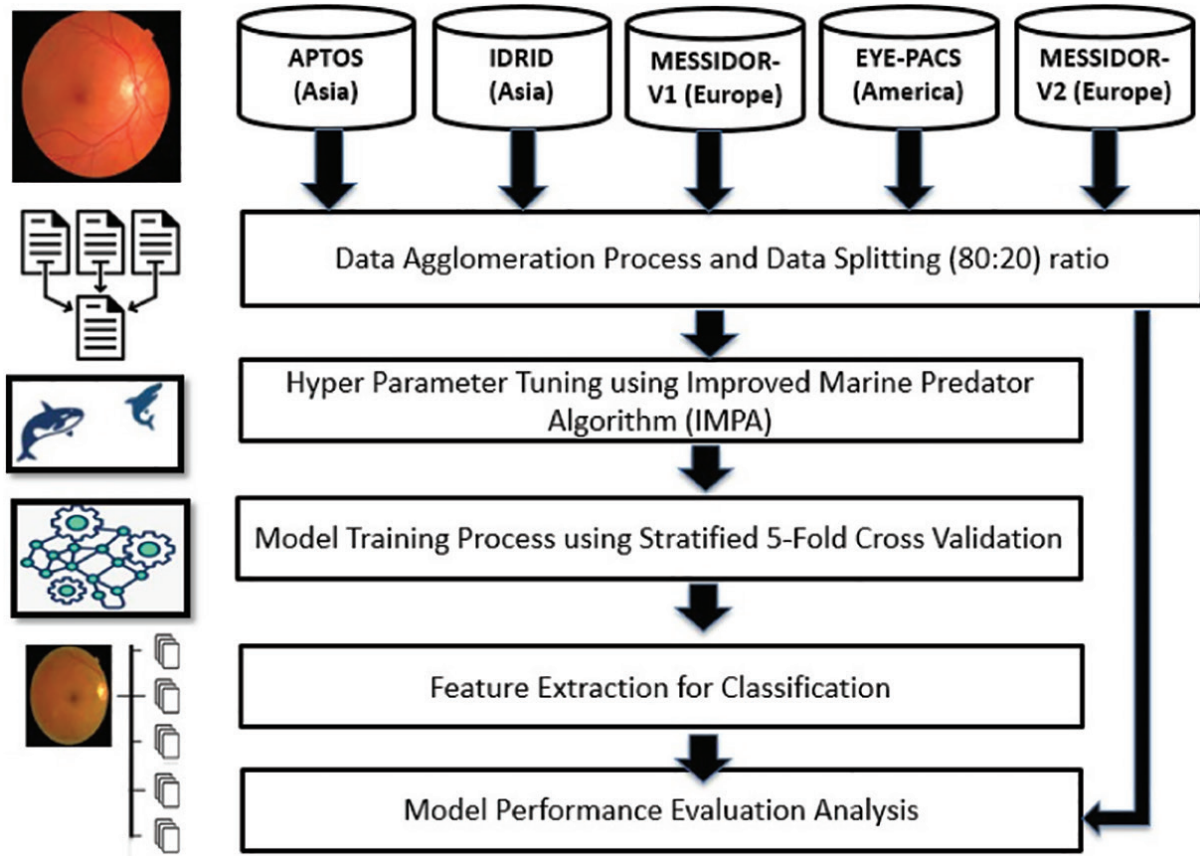


Figure 1: Proposed ARMDiaRD module-wise system flowchart adopted during experimentation. IMPA, improved marine predator algorithm.

color differentiation), brightness (global illumination), color saturation (intensity of color representation), and noise level (random pixel intensity variations).

These metrics were computed for each healthy image, allowing a quality-based ranking of the dataset. Down-sampling was then performed by selecting high-scoring images, thereby ensuring that the reduced dataset preserved clinically meaningful features.

a.i Data preprocessing

The preprocessing improves the input quality in the agglomeration pipeline as the data are captured from various settings with respect to light adjustments. The images were first converted to JPG format and later to the green channel to capture the depth of the images, and retinal structures are best represented in this spectrum due to the BGR loading convention in OpenCV. Each input RGB fundus image from the individual dataset is resized to a fixed resolution of 224×224 pixels and normalized to maintain a

consistent intensity distribution. Noise reduction was achieved by applying a Gaussian blur with a 5×5 kernel, followed by Contrast Limited Adaptive Histogram Equalization (CLAHE) to enhance local contrast and highlight subtle retinal features. The optional label unification ensures that the labels are consistent across datasets, simplifying the task of training a model on data from different sources. Then, CLAHE is employed to locally sharpen contrast and improve visibility of fine structures, particularly microaneurysms and blood vessels [43, 44]. To further improve model robustness and reduce overfitting, data augmentation was performed using the Albumentations library, which included transformations, such as horizontal flipping, brightness/contrast adjustment, random rotations, and Gaussian noise injection. Finally, the processed and augmented dataset was serialized into pickle format to allow efficient storage and rapid loading during model training. Later, the images are appended into the combined dataset and efficiently stored in a serialized format to facilitate faster data access during model training.

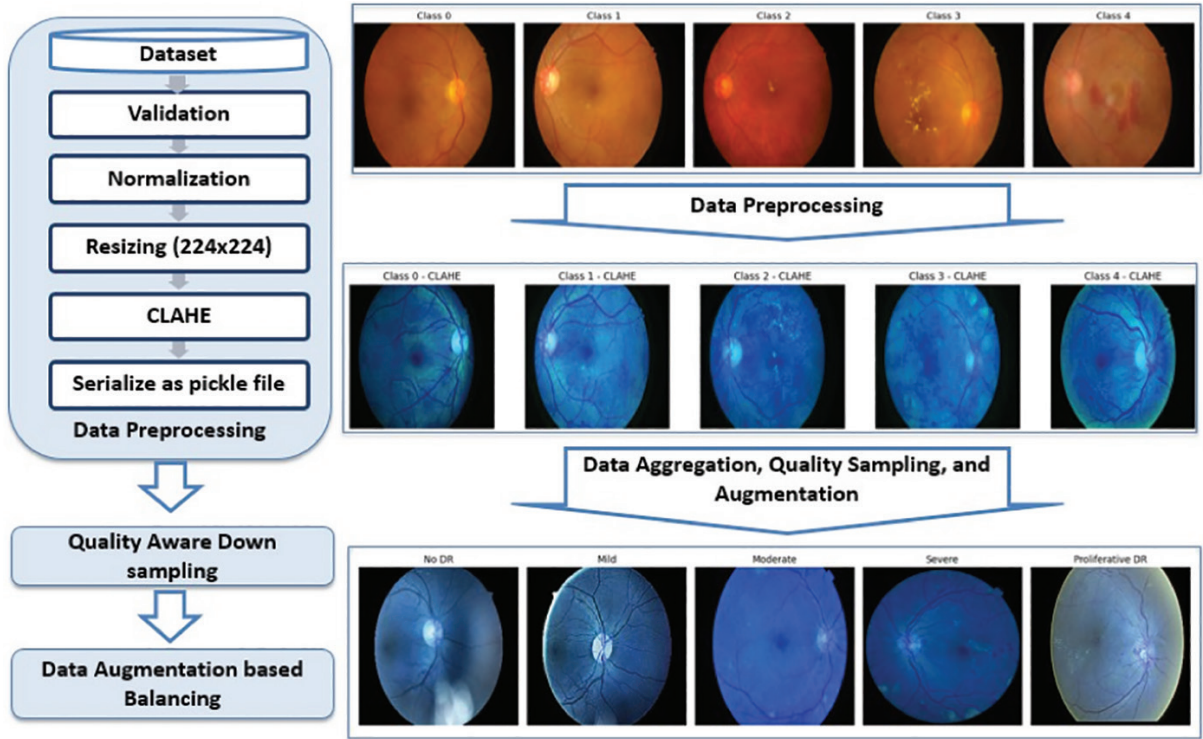


Figure 2: Data agglomeration pipeline with sample images showing the three stages of data preprocessing, quality-aware down-sampling, and augmentation ending with splitting for the training and testing process.

a.ii Quality-aware down-sampling

To ensure a balanced dataset suitable for training machine learning models, a quality-aware down-sampling strategy is employed on each dataset, as the fundus datasets are highly imbalanced. A threshold value of 7,000 images is set for the combined dataset for each class, consolidating to 35,000 images for all five classes. Each image is scored using a composite quality metric defined in Eq. (3), which combines sharpness, contrast, and brightness.

$$Q = 0.5 \cdot \text{Sharpness} + 0.3 \cdot \text{Contrast} + 0.2 \cdot \text{Brightness} \quad (3)$$

where Sharpness = variance of Laplacian [45],
 Contrast = standard deviation of pixel intensities.
 Brightness = mean pixel intensity.

Images with a low Q score are excluded from the training dataset. This filtering process improves the signal-to-noise ratio and ensures that the model focuses on diagnostically useful information. Quality-aware filtering is particularly beneficial when training architectures, such as Swin Transformers, which are sensitive to feature representation quality. The histogram results

shown in Figure 3 show how the quality of the original images varies with the down-sampled balanced images with respect to the Q -scores of the images selected for training in each class of DR grading. Table 3 shows the number of balanced images after the data agglomeration pipeline used for experimentation of the proposed system with its properties from the four datasets APTOS, IDRID, MESSIDOR V1&V2, and EYEPACS.

a.iii Data augmentation

To mitigate overfitting and enhance model generalization, each image undergoes stochastic data augmentation. Horizontal flip of probability 0.5, random adjustment of brightness and contrast of probability 0.2, and random rotation within $\pm 15^\circ$ of probability 0.5 are used to generate new images. Additionally, the sharpening using the Lambda transformation applies a custom sharpening filter to enhance image edges, improving clarity and enabling the model to detect finer details. These augmentations not only enrich the dataset by synthetically increasing its size but also expose the model to diverse variations in real-world scenarios.

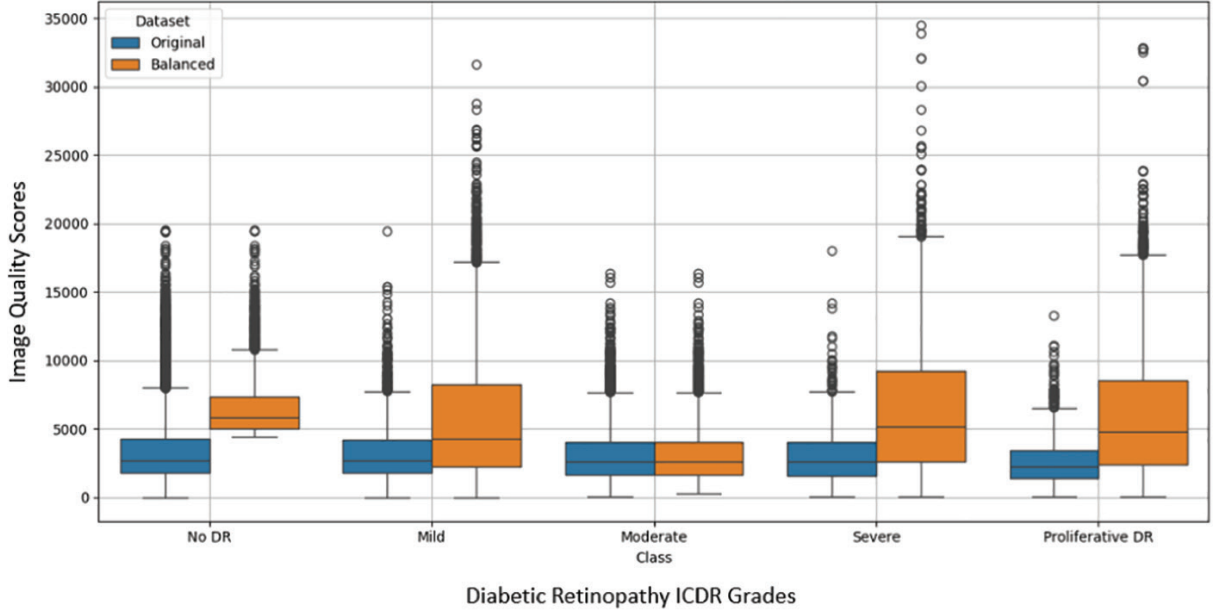


Figure 3: Histogram of Q-scores calculated for each grade of DR before and after quality-aware down-sampling, which reflects the improvement in quality of images selected for the oversampled classes. DR, diabetic retinopathy; ICDR, international clinical diabetic retinopathy.

b. Hyperparameter tuning using IMPO

In the proposed system ARMDiaRD, the IMPA-driven tuning process [41, 42] is integrated to evaluate each hyperparameter configuration reliably. The use of such optimization algorithms significantly accelerates the convergence toward an optimal set of HP and enhances the overall predictive performance of the model. The improved marine predators algorithm (IMPO) is used for hyperparameter optimization of the proposed hybrid Swin Transformer model with MSFF. The fitness function within IMPO evaluates the effectiveness of different hyperparameter combinations, including the learning rate (lr), the embedding dimension (embed dim), and the dropout rate (dr). The pseudocode of the IMPA optimizer process is shown in algorithm 1.

Algorithm 1 IMPA Hyperparameter Tuning with Stratified K-Fold Cross-Validation

Require: HP bounds $X_{\min}$, $X_{\max}$; folds k ; population size N ; max iterations T

Ensure: Best hyperparameter vector X^*

1: **Initialize:** Generate initial population X_0 containing lr, embed dim, dr of size N :

$$X_i = X_{\min} + \text{rand}() \times (X_{\max} - X_{\min}) \quad \forall i = 1, \dots, N$$

Example: $X_i = [\text{lr}, \text{dropout}, \text{embed dim}]$

2: **for** iteration $t = 1$ to T **do**

3: **for** each candidate solution X_i in population **do**

4: Initialize an empty list of fold scores

5: **for** fold $j = 1$ to k **do**

6: Split data using Stratified K-Fold to obtain training and validation sets

7: Train the model using X_i on training data of fold j

8: Evaluate the model on validation data of fold j

9: Record validation accuracy

10: **end for**

11: Compute average performance across k folds:

$$f(X_i) = \frac{1}{k} \sum_{j=1}^k \text{Accuracy}_j$$

$$\text{Fitness} = (1 - \text{Accuracy})$$

12: **end for**

13: Update population using 0.5 probability on exploration based on the best solution

14: **end for**

15: **Output:** Best hyperparameter vector X^* with minimal fitness

Table 3: Class-wise image summary after dataset agglomeration process used for ARMDiaDR system experimentation

textbfDataset	No DR	Mild NPDR	Moderate NPDR	Severe NPDR	PDR	Total Images	Dimensions
APTOS [20]	1,805	370	999	193	295	3,662	Mixed
IDRID [21]	168	25	168	93	62	516	4,288 × 2,848
EYEPACS [19]	25,802	2,438	5,288	872	708	35,108	Mixed
MESSIDOR- V1 [23]	546	153	247	254	–	1,200	2,240 × 1,488
MESSIDOR- V2 [46]	1,017	270	347	75	35	1,744	2,240 × 1,488
Training Dataset	5,600	5,600	5,600	5,600	5,600	28,000	224 × 224
Testing Dataset	1,400	1,400	1,400	1,400	1,400	7,000	224 × 224

DR, diabetic retinopathy; NPDR, non-proliferative diabetic retinopathy; PDR, proliferative diabetic retinopathy.

These bounds ensure that the optimizer explores a diverse and feasible region of the hyperparameter space. Along with this, the population size was set to 10, and the upper limit on the number of iterations was set to 5. The hyperparameter bounds used in the experimentation is as follows:

- **Learning Rate (lr):** The search interval was set to $[1 \times 10^{-5}, 1 \times 10^{-3}]$. This controls the step size in gradient-based updates.
- **Dropout Rate (dr):** The search interval was $[0.1, 0.5]$, which controls the fraction of neurons randomly dropped during training to prevent overfitting.
- **Embedding Dimension (embed dim):** This was searched within the range $[48, 192]$ and defines the dimensionality of intermediate feature representations in the model.

A stratified five-fold cross-validation split was used to maintain class distribution consistency. For each candidate solution X_i , the model is trained using the generated HP within these bounds. The trained model is then evaluated on a validation set to obtain a performance metric, specifically the validation accuracy for classification tasks. The fitness value is calculated as shown in the algorithm 1 with the minimizing problem. The best-performing hyperparameter configuration for the learning rate, dropout rate, and embedding dimension across all iterations and candidates is recorded and used for the proposed

ARMDiaRD system’s final training and testing process, as summarized in the results table.

c. Model architecture

The proposed ARMDiaRD system uses EfficientNet-B0, a baseline model, as its back-bone, initialized with ImageNet-1k pretrained weights. Built on it, a hybrid wiggled layers of Swin blocks with multiscale features fusion, as shown in Figure 4. The proposed system is a hybrid model that uses EfficientNet-B0, a lightweight and efficient CNN backbone, to extract initial image features as a baseline model. These features are enhanced using a GC block to capture broader spatial dependencies, followed by a 1×1 convolution layer for dimensional alignment. The core of the model consists of wiggled Swin Transformer blocks, each followed by a MSFF layer and global average pooling (GAP), enabling the model to capture both local and hierarchical global features. Outputs from each Swin block (GAP1, GAP2, and GAP3) are concatenated to form a rich feature representation, which is then passed to a fully connected layer for final classification.

c.i Baseline GC capture

A baseline model, EfficientNet-B0 backbone with a GC block, provides robust fundus image classification. The input image is first passed through the

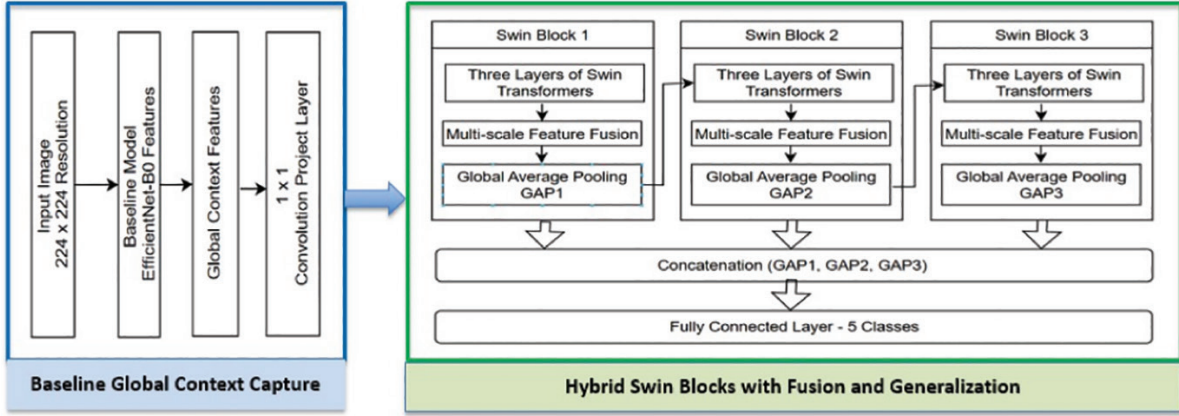


Figure 4: Proposed robust and hybrid deep learning model architecture with initially capturing the features using EfficientNet and GC block, followed by hybrid Swin Transformer blocks with MSFF and hierarchical aggregation. GC, global context; MSFF, multi-scale feature fusion.

EfficientNet-B0 backbone to extract rich hierarchical features while maintaining computational feasibility for high-resolution fundus images. The features generated by EfficientNet-B0 are then refined by a GC Block. Followed by with 1×1 convolutional layer is added to embed CNN features into transformer-friendly dimensions, minimizing computational overhead while enabling global information capture. For an input feature map $X \in \mathbb{R}^{B \times C \times H \times W}$, B is the batch size, C is the number of channels, and H and W are the spatial dimensions. The GC block computes its output Y as stated in the Eq. (4).

$$Y = X + W_v \left(\text{ReLU} \left(\text{LayerNorm} \left(W_k \left(X M^T \right) \right) \right) \right), \quad (4)$$

where, $M = \text{Softmax}(\text{Conv}_{1 \times 1}(X))$

The learnable weight matrices $W_k \in \mathbb{R}^{C \times C}$ and $W_v \in \mathbb{R}^{C \times C}$ is implemented as 1×1 convolutions that project the input feature to a lower-dimensional space and back, respectively. The final output Y adds the context to the original input with a residual connection. The GC block aggregates long-range dependencies by computing global attention maps, which help the model emphasize relevant regions of lesions, exudates, and microaneurysms that may be scattered across the retina. This enhances the model's robustness to variations in lesion size, shape, and location.

c.ii Swin blocks

Each Swin Transformer first normalizes the input, applies multi-head self-attention (MHSA) with a skip connection, normalizes again, then passes through a

feed-forward multilayer perceptron (MLP) with a skip connection. This residual connection helps gradients flow more effectively during backpropagation and prevents vanishing gradients, making deeper networks easier to train. It stacks three Swin Transformer blocks, and each block's output passes through a MSFF block consisting of convolution layers: Conv 1×1 , Conv 3×3 , and Conv 5×5 for feature fusion. GAP is performed individually on each stage, followed by flattening. Considering the input feature sequence X , the Swin Transformer Block applies two residual sublayers as stated in Eq. (5).

$$\begin{aligned} Z &= X + \text{MHSA}(\text{LayerNorm}_1(X)), \\ Y &= Z + \text{MLP}(\text{LayerNorm}_2(Z)), \end{aligned} \quad (5)$$

where LayerNorm denotes layer normalization, MHSA is the multi-head self-attention, and MLP is a feed-forward network that expands the channel dimension by a factor of 4 before projecting it back. The residual connections help to stabilize training and preserve input information.

c.iii Hybrid Swin with hierarchical fusion (HS-HF)

The robust hybrid ARMDiaRD uses three Swin blocks in a cascade that helps in learning progressively complex features at different levels. The outputs from all three blocks (low, mid, and high-level features) are aggregated hierarchically, by combining features at different levels of abstraction, the system learns a more comprehensive feature representation. A dropout layer is applied to the concatenated feature to prevent overfitting and improve

generalization. Finally, a linear classification head maps the aggregated representation to the targeted multi-classification problem or DR, having five classes as per ICDR grading. The final prediction is computed using a linear classification head applied to the concatenated multi-stage features using Eq. (6):

$$Y_{\text{pred}} = W_c Z + b \quad (6)$$

where Y_{pred} is the predicted class output, Z is the concatenated feature vector obtained from three Swin Transformer stages, W_c is the learnable classification weight matrix to the target classes, b is the learnable bias, and C denotes the number of output classes ($C = 5$).

IV. Implementation and Results

a. Experimental setup

For the experimental setup, a high-performance NVIDIA manufactured A100 80GB PCIe GPU, installed with PyTorch version 2.7.0 and CUDA version

12.2 from it's official website, is employed to ensure compatibility with the latest deep learning frameworks. During experimentation, the GPU operated at a stable temperature of 37°C with a power consumption of 67 W, well within its maximum capacity of 300 W.

b. Evaluation metrics

To evaluate the proposed system ARMDiaRD, the evaluation metrics as listed in Table 4, with their importance and formulas, are adopted. Along with the normal metrics, such as accuracy, recall, precision, F1-score, the class-wise performance metric, QWK, and Spearman correlation, which considers the deviation from the original class in ordinal prediction problems, such as DR detection from no DR (Class 0) to PDR (Class 4), is also calculated. The class differences are measured by how far the false predictions fall with this metric, and QWK and Spearman correlation play a vital role in the proposed model evaluation.

Table 4: Evaluation performance metrics used

Performance Metric	Mathematical Expression	Description
Accuracy	$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total Instances}}$	Provides a high-level overview of model performance.
Recall	$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$	Measures how many actual positive instances the model correctly identified.
Precision	$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$	Measures the accuracy of positive predictions.
F1-Score	$\text{F1-Score} = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{\text{Precision} + \text{Recall}}$	Harmonic mean of precision and recall; balances false positives and false negatives.
QWK	$K = \frac{\text{Observed} - \text{Expected}}{1 - \text{Expected}}$	Measures agreement between two raters, penalizing bigger disagreements quadratically.
MAE	$\text{MAE} = \frac{1}{n} \sum_{i=1}^n Y_i - \hat{Y}_i $	Measures average magnitude of the errors between predicted and actual values.
MSE	$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	Similar to MAE but penalizes larger errors more by squaring them.
Spearman Correlation	$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$	Measures strength and direction of a monotonic relationship in ranked or ordinal data.

MAE, mean absolute error; MSE, mean squared error; QWK, Quadratic Weighted Kappa.

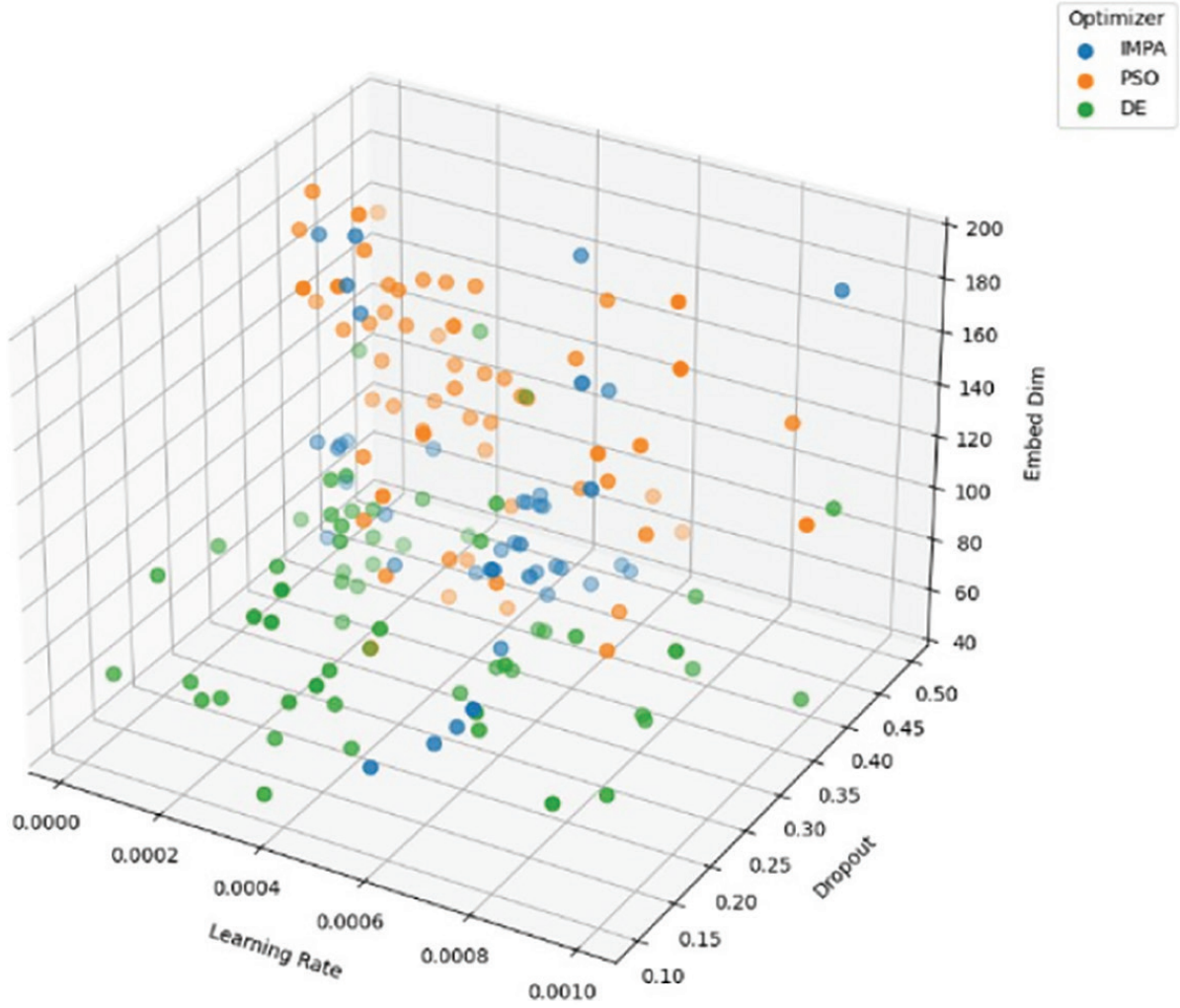


Figure 5: Scatter plot of the HP lr , dr , and $embed\ dim$ for the three optimizer algorithms, IMPO, PSO and DE. DE, differential evolution; dr , dropout rate; HP, hyperparameters; lr , learning rate; PSO, particle swarm optimizer.

c. Hyperparameter evaluation

The hyperparameter tuning is performed for the experiment by a metaheuristic optimizer, the improved marine predator optimization (IMPO) algorithm, by training on the entire dataset for 10 epochs. Initially, the entire dataset is subjected to three nature-inspired optimizers, the improved marine predator optimization (IMPO) algorithm, the particle swarm optimizer (PSO), and the differential evolution (DE) to find the best optimizer for the given problem. The pair plot shown in Figure 5 visualizes the distribution of the three HP, learning rate lr , dropout dr , and embedding dimension $embed\ dim$ explored during the exploration and exploitation process of the optimizers IMPO, PSO, and DE.

The one-way analysis of variance (ANOVA) test is performed to determine the best optimizer based

on accuracy, resulting in an F-statistic of 0.102 and a p -value of 0.904, indicating that there is no statistically significant difference among the three. However, IMPO is selected for intense hyperparameter tuning for the proposed system due to its nature of exploring a diverse range of options and achieving high accuracy results, as shown in Figure 6. Thus, the best optimized HP given by the IMPO optimizer fitness score listed in Table 5 are used for the training process.

d. Model training

Five publicly available datasets, EyePACS, APTOS, MESSIDOR-V1, MESSIDOR-V2, and IDRiD, is used for training and validation analyses. The preprocessed, quality-aware, and augmented datasets

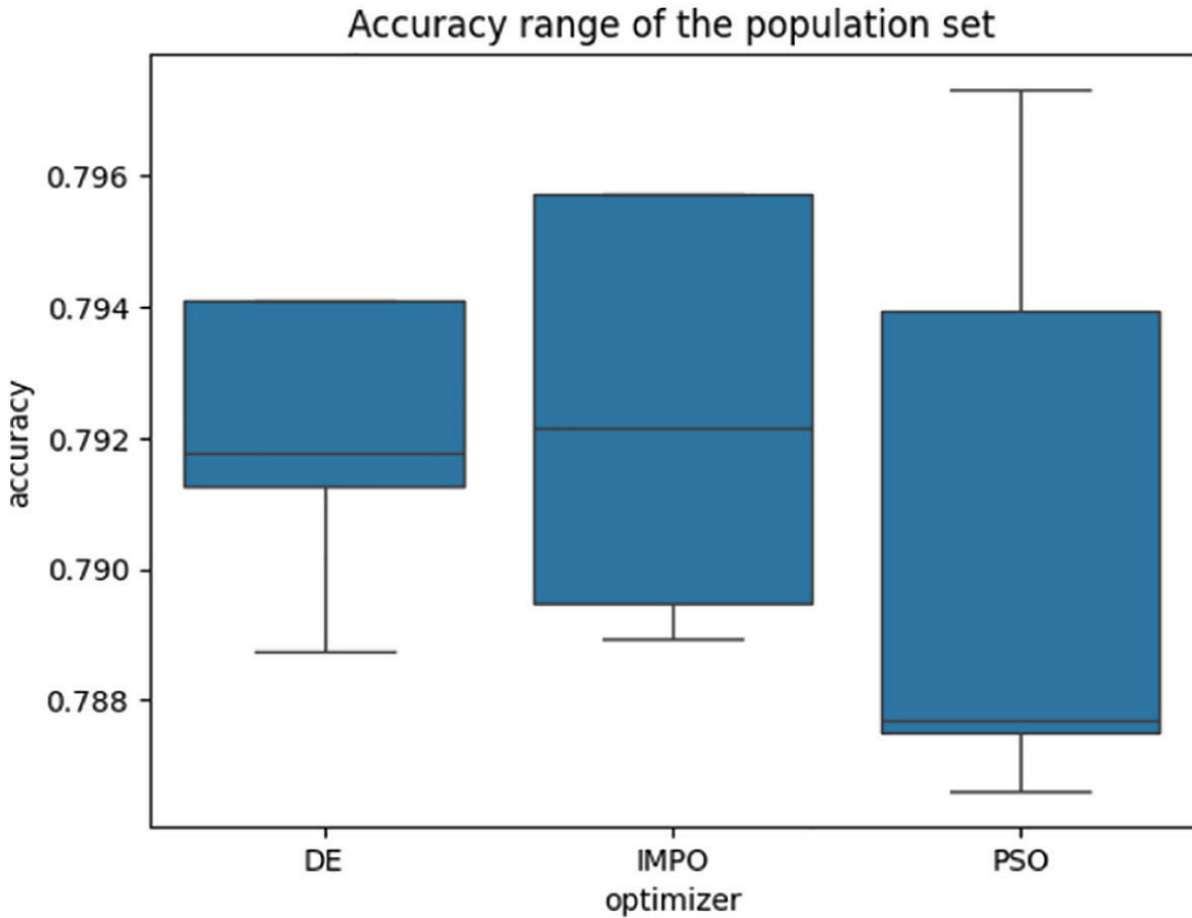


Figure 6: Accuracy box plot obtained from the three optimizer algorithms, DE, IMPO, and PSO. DE, differential evolution; PSO, particle swarm optimizer.

Table 5: HP achieved using IMPO optimizer for training the proposed ARMDiaRD

Sl. No.	Hyperparameter	Specification
01	Learning rate	0.00039
02	Number of epochs	40
03	Batch size	32
04	Dropout	0.428
05	Embed dim	80

HP, hyperparameters; IMPO, improved marine predator optimization.

comprise a total of 35,000 images is used in training, validation, and testing. The raw images were scaled to a resolution of 224×224 pixels and then normalized, and the agglomeration process was carried out as mentioned in Section 3.1 to improve the visibility of the lesion. The augmented dataset produced from the five various sources after the intense data

agglomeration process was divided using an 80%–20% division for training and evaluation. To ensure reproducibility, we fixed the random seed across NumPy, PyTorch, and Python built-in libraries. By stratifying the folds, the five-fold cross-validation helps preserve the class balance in every fold, leading to a more reliable estimation of the generalization

performance of the model. The generalized dataset used was trained and tested across various existing baseline classifier models listed in the cases below to validate the proposed system.

- *Case 1: ResNet*, the ResNet architecture of the baseline model (ResNet50) without any attention or transformer modules.
- *Case 2: EfficientNet-B0 (EffiNet)*, the lightweight and efficient EfficientNet-B0 backbone, pretrained on ImageNet without any attention or transformer modules.
- *Case 3: EffiNet with Global Context (EffiNet + GC)*, for contextual understanding, the GC block is added to the EfficientNet-B0 backbone.
- *Case 4: EffiNet + GC and Squeeze-and-Excitation (EffiNet + GC + SE)*, building on the previous configuration (EffiNet + GC), introduced Squeeze-and-Excitation (SE) blocks for channel-wise attention. SE blocks adaptively recalibrate feature maps to emphasize informative features, thereby improving representational power.
- *Case 5: EffiNet + GC and CBAM (EffiNet + GC + CBAM)*, this variant replaces SE, incorporating the Convolutional Block Attention Module (CBAM), which integrates both channel-wise and spatial attention.

- *Case 6: EffiNet + GC with 3 Swin Transformer Blocks and Multiscale Feature Fusion (Proposed)*, the final model integrates three sequential Swin Transformer blocks following the EfficientNet backbone. Their outputs are fused using a MSFF mechanism, enabling the model to combine hierarchical transformer features adaptively across scales.

The comparison line chart shown in Figure 7 clearly shows that the proposed system has better performance compared to the various other models on the same dataset processed by the proposed system. The accuracy, precision, recall, and F1 score obtained during the training process in the five folds for the 40 epochs for the different cases are shown in Figures 8–12, with the proposed system in Figure 13. The proposed system has performed best in its training process with a weighted average accuracy of 87.42%, recall of 87.49%, and precision of 87.51% among all the cases mentioned.

e. Model testing and ablation analysis

The testing model for the proposed system ARMDiaRD is developed using an ensemble evaluation of all five models generated through the five-fold

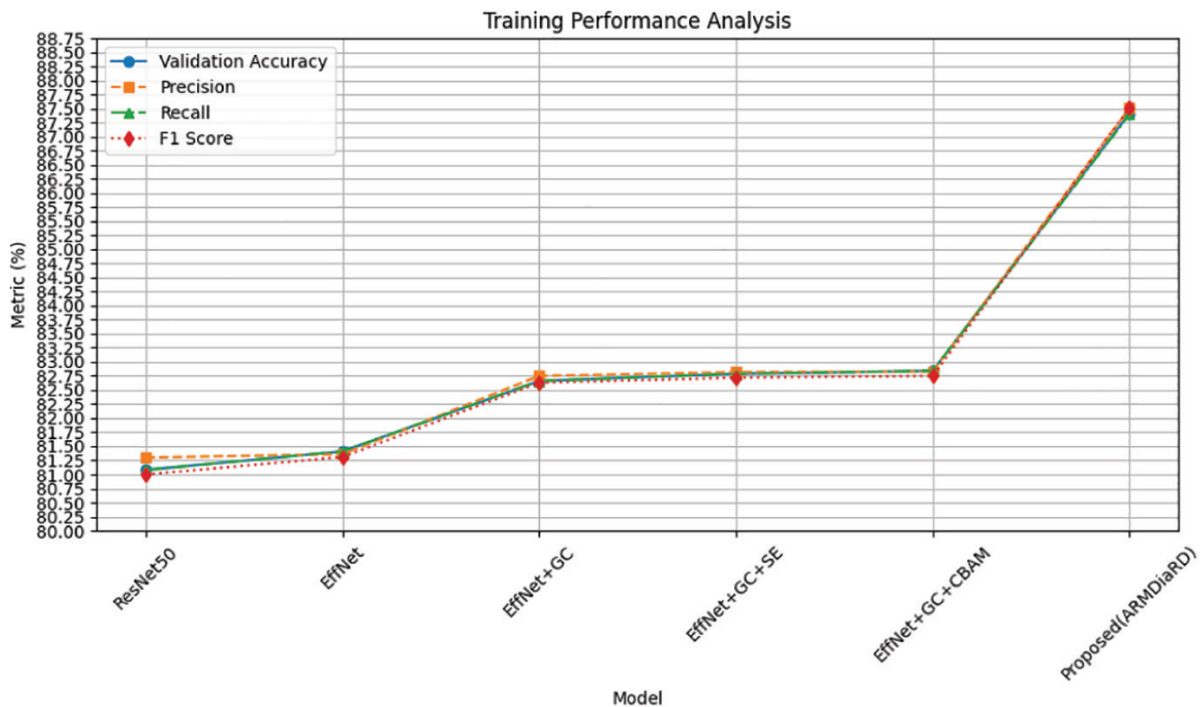


Figure 7: Results from stratified five-fold validation performed with average metrics of accuracy, precision, recall, and F1-score. CBAM, convolutional block attention module; GC, global context.

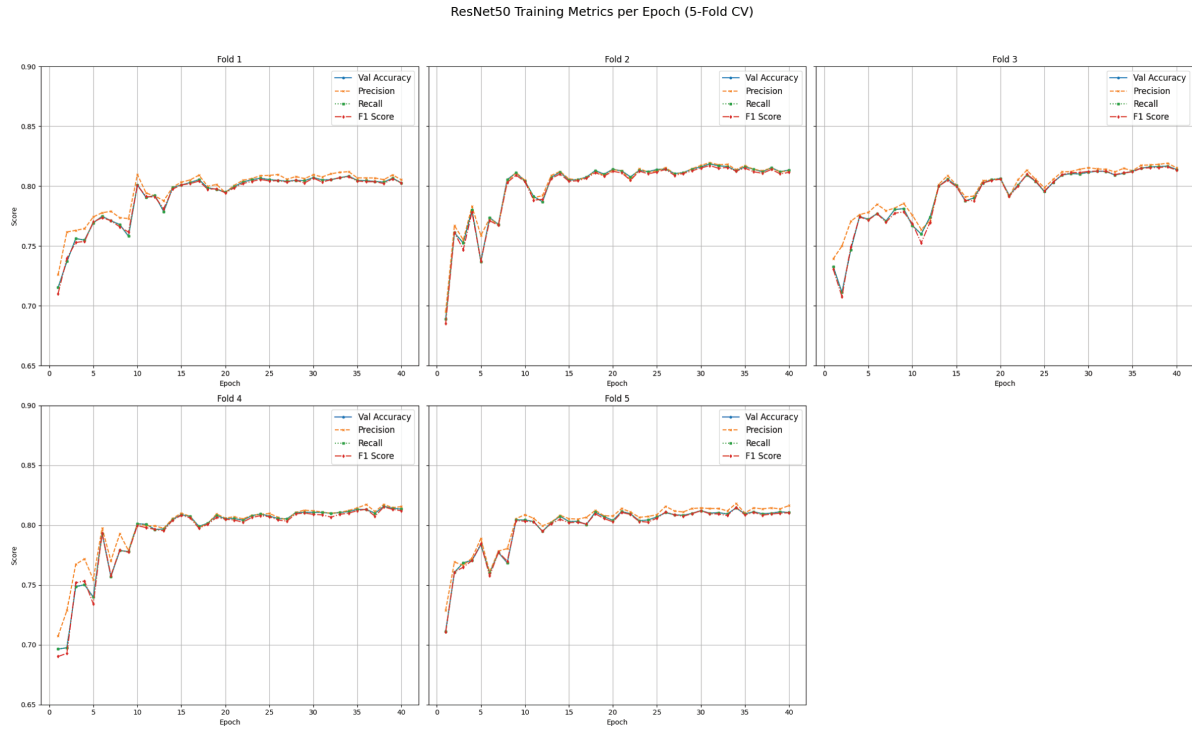


Figure 8: Results from stratified five-fold validation performed during the training phase using Resnet-50.

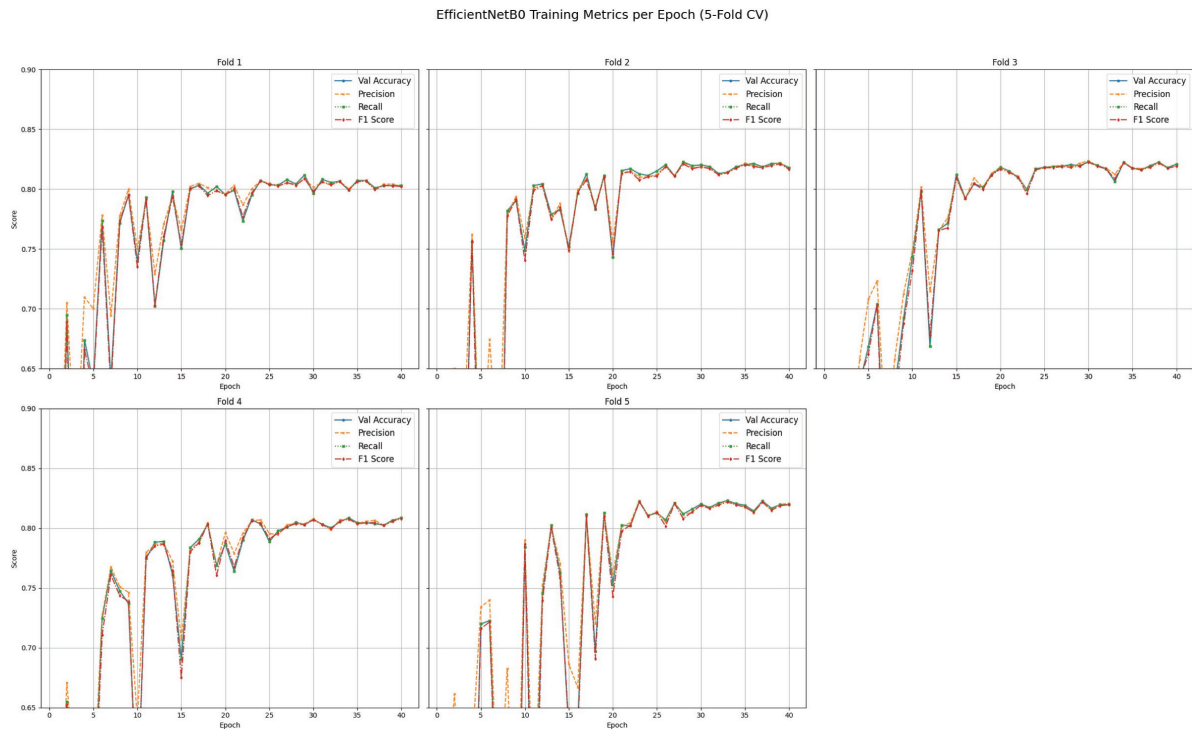


Figure 9: Results from stratified five-fold validation performed during the training phase using EfficientNet-B0.

EfficientNetB0_GC Training Metrics per Epoch (5-Fold CV)

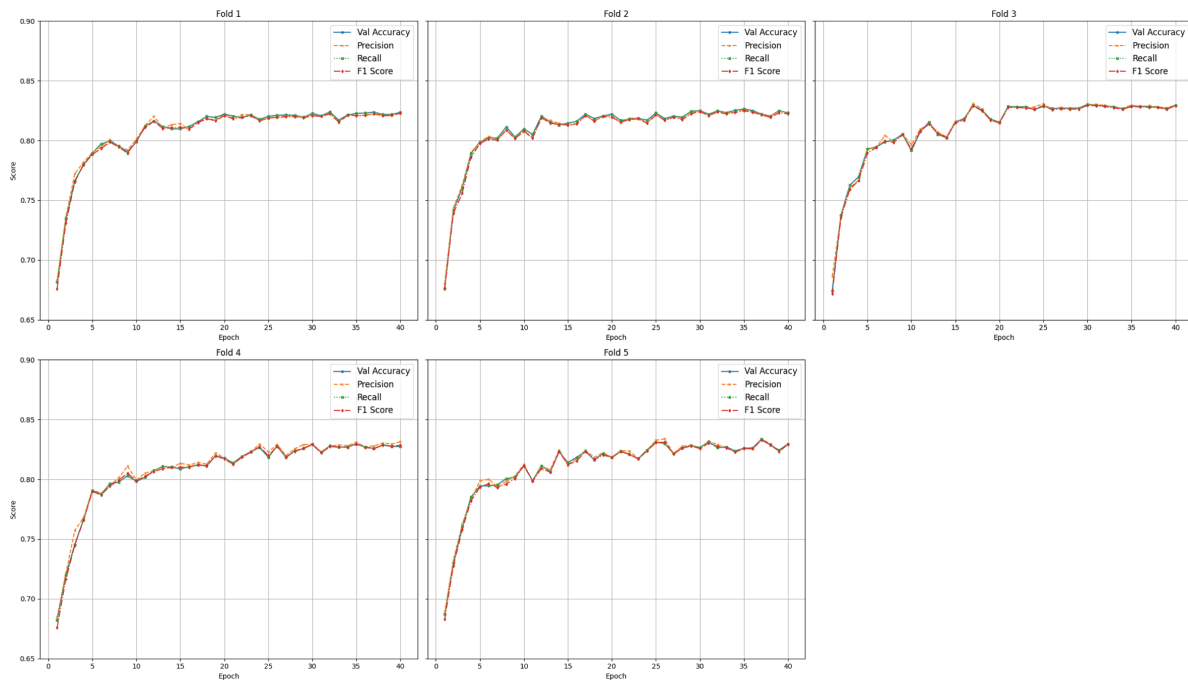


Figure 10: Results from stratified five-fold validation performed during the training phase using EfficientNet-B0 with GC. GC, global context.

EfficientNetB0_GC_SE Training Metrics per Epoch (5-Fold CV)

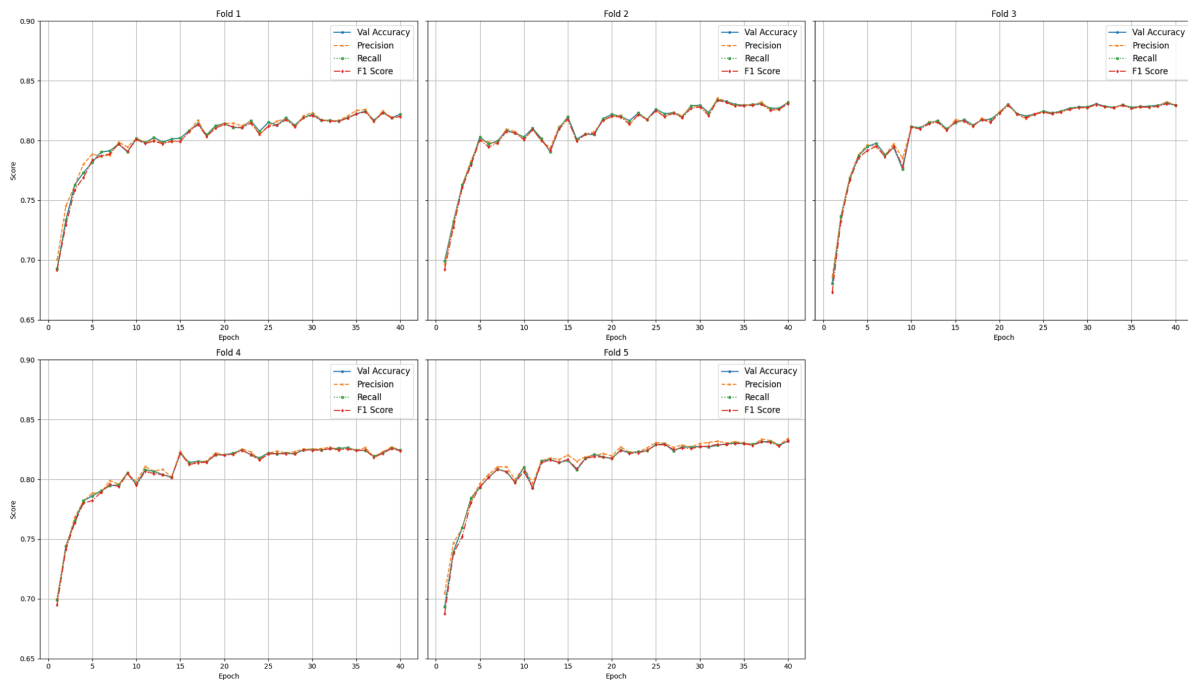


Figure 11: Results from stratified five-fold validation performed during the training phase using EfficientNet-B0 + GC + SE. GC, global context.

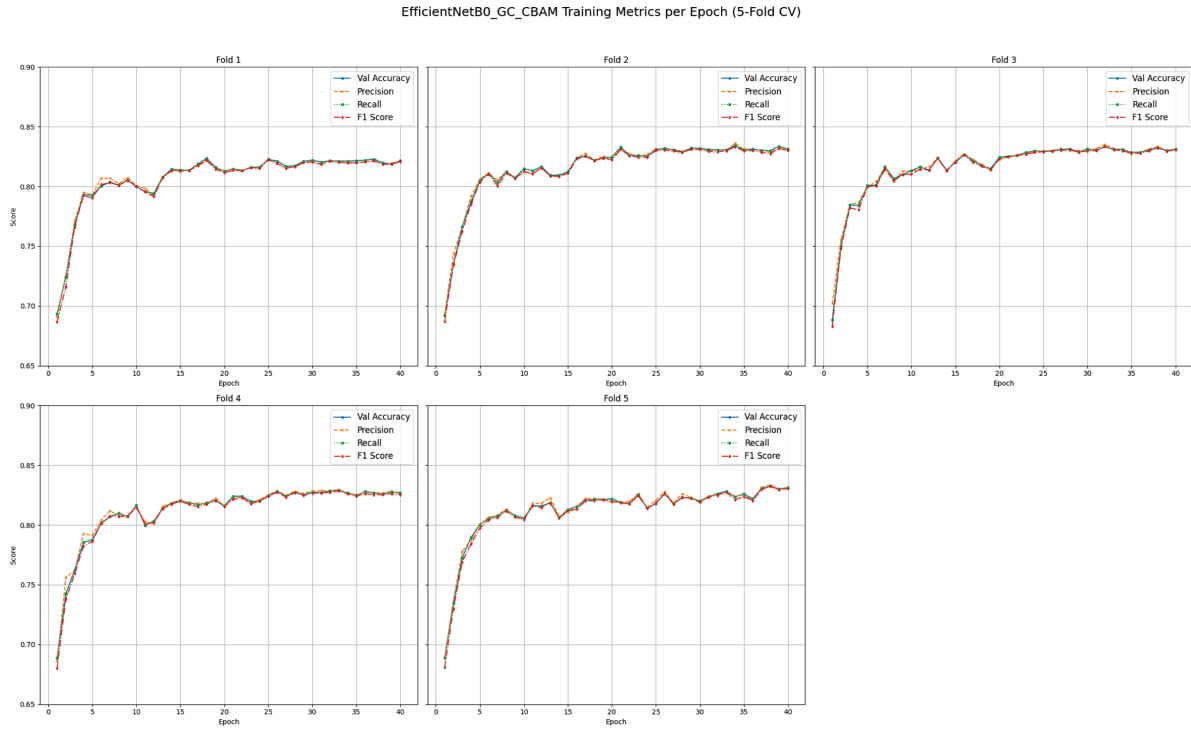


Figure 12: Results from stratified five-fold validation performed during the training phase using EfficientNet-B0 + GC + CBAM. CBAM, convolutional block attention module; GC, global context.

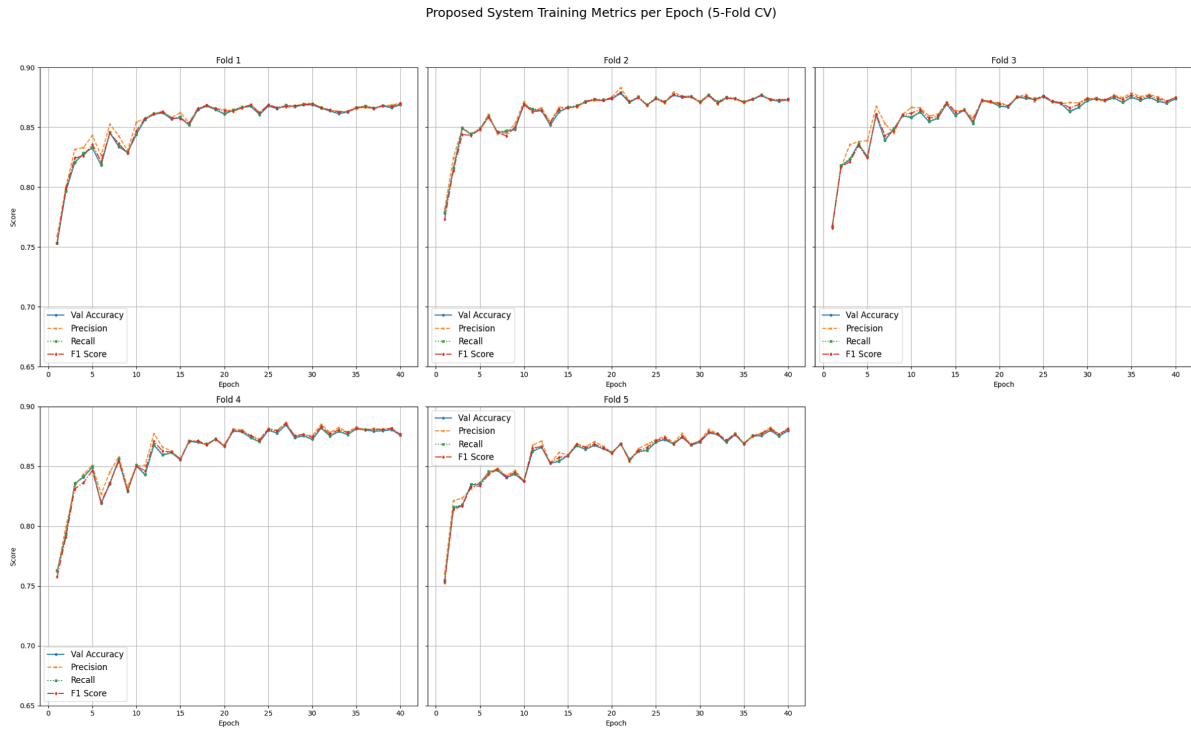


Figure 13: Results from stratified five-fold validation performed during the training phase using the proposed model ARMDiaRD.

cross-validation training phase. The model achieved an accuracy of approximately 87.59%, with a precision of about 87.6% and a recall of around 87.9%. The QWK and Spearman coefficients, which help in analyzing the variation from the actual score in case of wrong prediction, are high, with 0.9147 and 0.9253. This shows the proposed hybrid Swin layers capture context at multiple scales and complement the GC with a channel attention pipeline. The important metric in medical diagnosis of false predictions calculated by the mean absolute error (MAE) drops from 0.3310 to 0.1736, reflecting a reduction in prediction error magnitude in the proposed system by adding a hybrid Swin Transformer with hierarchical fusion of features. Importantly, the MAE, a critical metric for evaluating false predictions in medical diagnosis, dropped significantly from 0.3310 (baseline) to 0.1736 in the proposed model. However, the integration of a hybrid Swin Transformer on top of the baseline model, combined with training on a large dataset, has ultimately outperformed all other configurations. The graph shown in Figures 14 and 15 gives the ablation study results with clear advantages in both classification accuracy and ordinal consistency (QWK, Spearman), while also minimizing the MAE. Overall, the proposed hybrid ARMDiaRD system integrating GC and channel attention into a Swin Transformer

achieves the highest performance across all key metrics: accuracy, precision, recall, QWK, and Spearman correlation, along with the lowest MAE. These findings confirm that combining local, global, and hierarchical features within a carefully designed hybrid architecture is highly effective for robust medical image analysis.

V. Conclusion

In this research, the model is designed to handle robust multiclass image classification tasks leveraging the combined use of EfficientNet for effective feature extraction and a Swin Transformer to capture global contextual information, and multiscale feature fusion to enhance the model's ability to learn from global and local contexts. This aspect is significant for tasks involving complex visual patterns, where understanding both fine details and broader patterns is essential. Notably, compared with various existing DR detection models trained on large datasets, the proposed ARM-DiaRD hybrid Swin Transformer system with hierarchical feature fusion achieves the best overall results across multiple metrics, particularly in terms of QWK, MAE, and Spearman coefficients, which indicate the deviation in cases of incorrect predictions. Another major challenge in the field of DR datasets

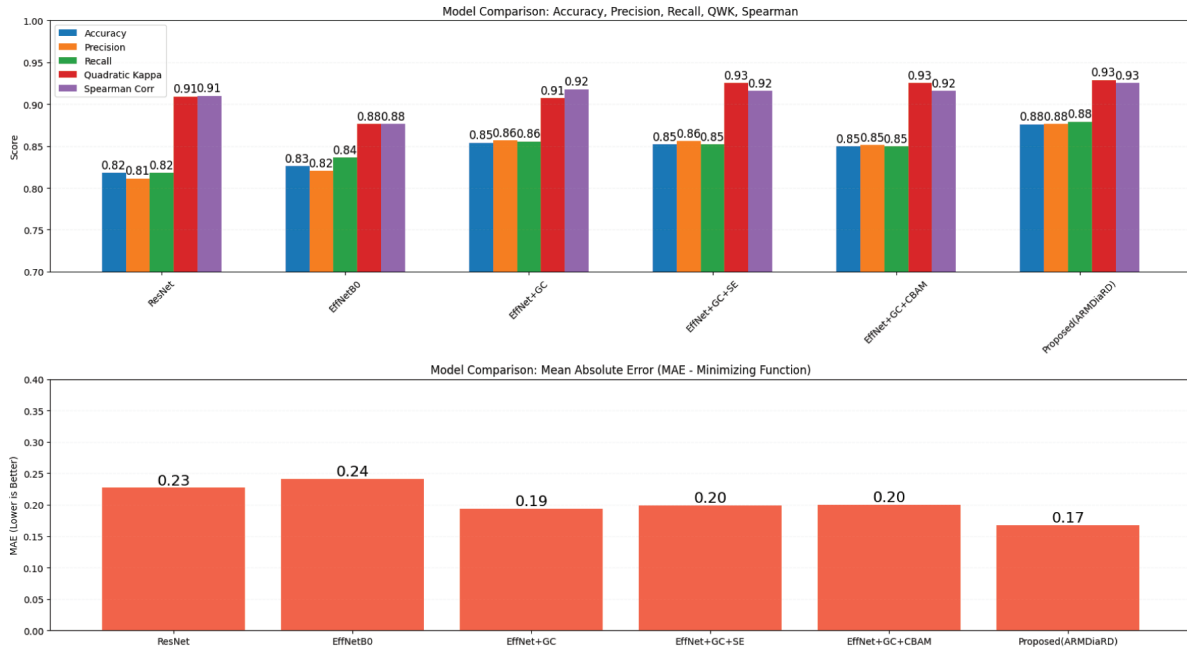


Figure 14: Results of ablation analysis performed with the various metrics, such as accuracy, precision, recall, QWK, Spearman correlation, and MAE with the training dataset subjected to the proposed agglomeration process, showing the overall improvement of the proposed model. MAE, mean absolute error; QWK, quadratic weighted kappa.

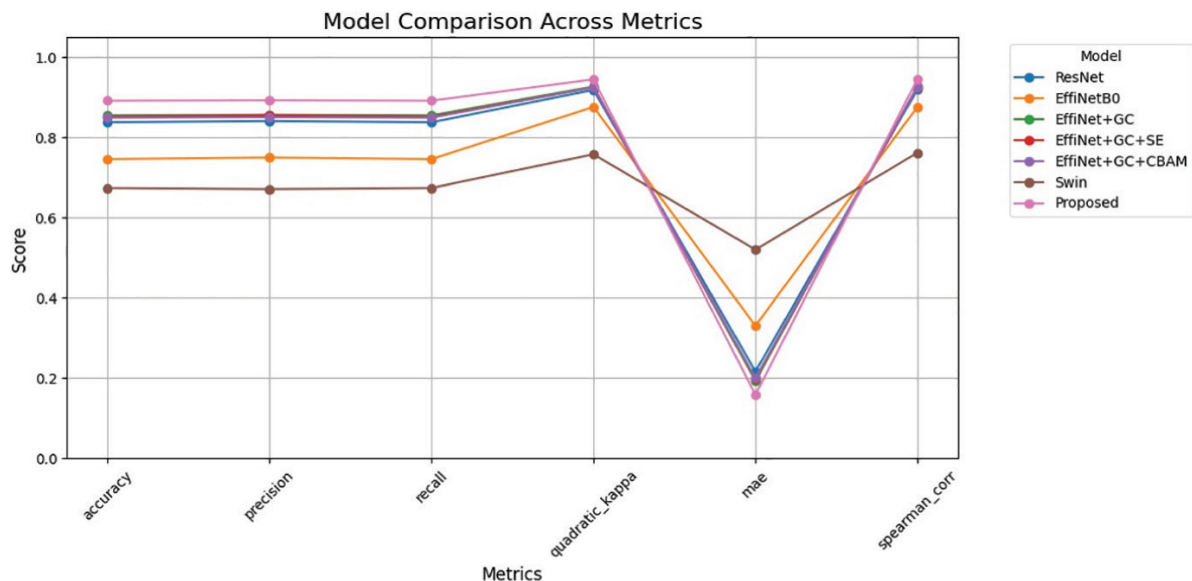


Figure 15: Results obtained during the testing process with metrics, such as accuracy, precision, recall, QWK, Spearman correlation, and MAE. MAE, mean absolute error; QWK, quadratic weighted kappa.

is severe class imbalance, which is addressed in this work through an intensive quality-aware agglomeration of five established ICDR-graded benchmark datasets. Despite the strong performance of the proposed hybrid model, certain limitations exist. Transformer-based architectures combined with MSFF can sometimes limit system adaptability compared to purely convolutional models, which naturally capture hierarchical feature flows and may generalize more easily to unseen variations. The future work can also be carried out by integrating segmentation to highlight the lesion regions, thereby increasing the performance of the system. Additionally, while the model demonstrates robust performance on aggregated benchmark datasets, validation on real-world patient data with ophthalmologist annotations would provide crucial external verification and further establish clinical reliability.

Declarations

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Author contribution

J Dhiviya Rose contributed to the main manuscript under the guidance of Ved Prakash Bhardwaj in material preparation, data collection, and analysis. J Dhiviya Rose wrote the first draft of the manuscript, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Ethics approval

Not applicable.

Consent to participate

The study does not involve the participation of human subjects. Therefore, the need for traditional informed consent is not applicable.

Consent to publish

Not applicable.

References

- [1] Li, T., *et al.*: Applications of deep learning in fundus images: A review. Elsevier B.V. (2021) DOI: 10.1016/j.media.2021.101971
- [2] Ting, D.S.W., Cheung, G.C.M., Wong, T.Y.: Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. Blackwell Publishing (2016) DOI: 10.1111/ceo.12696
- [3] Nagpal, D., Panda, S.N., Malarvel, M., Pattanaik, P.A., Khan, M.Z.: A review of diabetic retinopathy: Datasets, approaches, evaluation metrics and future trends. King Saud bin Abdulaziz University (2022). DOI: 10.1016/j.jksuci. 2021.06.006
- [4] Tang, L., Xu, G.T., Zhang, J.F.: Inflammation in diabetic retinopathy: possible roles in pathogenesis and potential implications for therapy. Wolters Kluwer Medknow Publications (2023) DOI: 10.4103/1673-5374.355743
- [5] Yau, J.W.Y., *et al.*: Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care 35(3), 556–564 (2012) DOI: 10.2337/dc11-1909
- [6] Abramoff, M.D., Folk, J.C., Han, D.P., *al.*: Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA Ophthalmology 131(3), 351–357 (2013) DOI: 10.1001/jamaophthalmol.2013.1743
- [7] Bora, A., Balasubramanian, S., Babenko, B., Virmani, S., Venugopalan, S., Mitani, A., Oliveira Marinho, G., Cuadros, J., Ruamviboonsuk, P., Corrado, G.S., Peng, L., Webster, D.R., Varadarajan, A.V., Hammel, N., Liu, Y., Bavishi, P.: Predicting the risk of developing diabetic retinopathy using deep learning. The Lancet Digital Health 3, 10–19 (2021) DOI: 10.1016/S2589-7500(20)30250-8
- [8] Serener, A., Serte, S.: Geographic variation and ethnicity in diabetic retinopathy detection via deep learning. Turkish Journal of Electrical Engineering and Computer Sciences 28, 664–678 (2020) DOI: 10.3906/elk-1902-131
- [9] Wilkinson, C.P., Ferris, F.L., Klein, R.E., *et al.*: Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology 110(9), 1677–1682 (2003)
- [10] Lakshminarayanan, V., Kheradfallah, H., Sarkar, A., Balaji, J.J.: Automated detection and diagnosis of diabetic retinopathy: A comprehensive survey. MDPI (2021). DOI: 10.3390/jimaging7090165
- [11] Ai, Z., Huang, X., Fan, Y., Feng, J., Zeng, F., Lu, Y.: Driixrn: Detection algorithm of diabetic retinopathy based on deep ensemble learning and attention mechanism. Frontiers in Neuroinformatics 15 (2021) DOI: 10.3389/fninf.2021.778552
- [12] Dhiviya Rose, J., Jain, A., Tiwari, S.: Challenges and solutions with lightweight models for diabetic retinopathy detection. In: International Conference on MAchine inTElligence for Research & Innovations, pp. 95–105 (2023). Springer
- [13] Mutawa, A., Alnajdi, S., Sruthi, S.: Transfer learning for diabetic retinopathy detection: A study of dataset combination and model performance. Applied Sciences 13(9), 5685 (2023)
- [14] Ansari, G.A., Bhat, S.S., Ansari, M.D.: Machine learning techniques for diabetes mellitus based on lifestyle predictors. Recent Advances in Electrical & Electronic Engineering 18(7), 1060–1071 (2025)
- [15] Bhat, S.S., Ansari, G.A., Ansari, M.D.: Performance analysis of machine learning based on optimized feature selection for type ii diabetes mellitus. Multimedia Tools and Applications 84(8), 4945–4964 (2025)
- [16] Bhat, S.S., Selvam, V., Ansari, G.A., Ansari, M.D.: Hybrid prediction model for type-2 diabetes mellitus using machine learning approach. In: 2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC), pp. 150–155 (2022). IEEE
- [17] Bhat, S.S., Selvam, V., Ansari, G.A., Ansari, M.D., Rahman, M.H.: Prevalence and early prediction of diabetes using machine learning in north kashmir: a case study of district bandipora. Computational Intelligence and Neuroscience 2022(1), 2789760 (2022)
- [18] Bhat, S.S., Selvam, V., Ansari, G.A., Ansari, M.D.: Analysis of diabetes mellitus using machine learning techniques. In: 2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT), pp. 1–5 (2022). IEEE
- [19] Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., *et al.*: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316(22), 2402–2410 (2016)
- [20] Community, K.: APTOS 2019 Blindness Detection. <https://www.kaggle.com/competitions/aptos2019-blindness-detection> (2019)
- [21] Porwal, P., Pachade, S., Kamble, R., *et al.*: Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research. Data 3(3), 25 (2018)
- [22] Jiang, Z., Fu, H., Shen, J., Xu, D., Shao, L.: Fgadr: Fine-grained annotated diabetic retinopathy dataset. arXiv preprint arXiv:2004.06803 (2021)
- [23] Decenciere, E., Zhang, X., Cazuguel, G., *et al.*: Feedback on a publicly distributed image database:

the messidor database. *Image Analysis & Stereology* 33(3), 231–234 (2014)

[24] Decenciere, E., Zhang, X., Cazuguel, G., *et al.*: Teleophtha: Machine learning and image processing methods for teleophthalmology. *IRBM* 34(2), 196–203 (2013)

[25] Baba, S.M., Bala, I., Dhiman, G., Sharma, A., Viriyasitavat, W.: Automated diabetic retinopathy severity grading using novel dr-resnet+ deep learning model. *Multimedia Tools and Applications* 83(28), 71789–71831 (2024)

[26] Islam, M. M., Yang, H.-C., Poly, T. N., Jian, W.-S., & Li, Y.-C. J. (2020). *Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis*. *Computer Methods and Programs in Biomedicine*, 191, 105320. <https://doi.org/10.1016/j.cmpb.2020.105320>.

[27] 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>

[28] Wang, X., Ji, Z., Ma, X., Zhang, Z., Yi, Z., Zheng, H., Fan, W., & Chen, C. (2021). *Automated grading of diabetic retinopathy with ultra-widefield fluorescein angiography and deep learning*. *Journal of Diabetes Research*, 2021, Article 2611250. <https://doi.org/10.1155/2021/2611250>.

[29] Voets, M., Möllersen, K., & Bongo, L. A. (2019). *Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs*. *PLoS ONE*, 14(6), e0217541. <https://doi.org/10.1371/journal.pone.0217541>.

[30] Tiwari, S., Jain, A., Ahuja, N.J., Shukla, A.: Deep learning-based multi-class classification of diabetic retinopathy utilizing transfer learning with mobilenet architecture. In: Rathore, V.S., Piuri, V., Babo, R., Tiwari, V. (eds.) *Emerging Trends in Expert Applications and Security*, pp. 83–92. Springer, Singapore (2024)

[31] Lin, C.-L., Wu, K.-C.: Development of revised resnet-50 for diabetic retinopathy detection. *BMC bioinformatics* 24(1), 157 (2023)

[32] Al-Antary, M.T., Arafa, Y.: Multi-scale attention network for diabetic retinopathy classification. *IEEE Access* 9, 54190–54200 (2021) DOI: 10.1109/ACCESS.2021.3070685

[33] Abramovich, O., Pizem, H., Van Eijgen, J., Oren, I., Melamed, J., Stalmans, I., Blumenthal, E.Z., Behar, J.A.: Fundusq-net: A regression quality assessment deep learning algorithm for fundus images quality grading. *Computer methods and programs in biomedicine* 239, 107522 (2023)

[34] Koonce, B.: Resnet 50. In: *Convolutional Neural Networks with Swift for Ten-sorflow: Image Recognition and Dataset Categorization*, pp. 63–72. Springer, (2021)

[35] Alwakid, G., Gouda, W., Humayun, M.: Deep learning-based prediction of diabetic retinopathy using clahe and esrgan for enhancement. In: *Healthcare*, vol. 11, p. 863 (2023). MDPI

[36] Mohanty, C., Mahapatra, S., Acharya, B., Kokkoras, F., Gerogiannis, V.C., Karamitsos, I., Kanavos, A.: Using deep learning architectures for detection and classification of diabetic retinopathy. *Sensors* 23(12), 5726 (2023)

[37] Deepak, G.D., Bhat, S.K.: Deep learning-based CNN for multiclassification of ocular diseases using transfer learning. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 12(1), 2335959 (2024)

[38] Dai, L., Sheng, B., Chen, T., Wu, Q., Liu, R., Cai, C., Wu, L., Yang, D., Hamzah, H., Liu, Y., *et al.*: A deep learning system for predicting time to progression of diabetic retinopathy. *Nature Medicine* 30(2), 584–594 (2024)

[39] Al-Naji, A., Khalid, G.A., Mahmood, M.F., Chahl, J.: Computer vision for eye diseases detection using pre-trained deep learning techniques and raspberry pi. *The Journal of Engineering* 2024(7), 12410 (2024)

[40] Gupta, I.K., Patil, S., Mahadevkar, S., Kotecha, K., Mishra, A.K., Rodrigues, J.J.: Retinal fundus imaging-based diabetic retinopathy classification using transfer learning and fennec fox optimization. *MethodsX* 14, 103232 (2025)

[41] Faramarzi, A., Heidarinejad, M., Mirjalili, S., Gandomi, A.H.: Marine predators algorithm: A nature-inspired metaheuristic. *Expert systems with applications* 152, 113377 (2020)

[42] Chun, Y., Hua, X., Qi, C., Yao, Y.X.: Improved marine predators algorithm for engineering design optimization problems. *Scientific reports* 14(1), 13000 (2024)

[43] Stimper, V., Bauer, S., Ernstorfer, R., Schölkopf, B., Xian, R.P.: Multidimensional contrast limited adaptive histogram equalization. *IEEE Access* 7, 165437–165447 (2019)

[44] Reza, A.M.: Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology* 38, 35–44 (2004)

[45] Pertuz, S., Puig, D., Garcia, M.A.: Analysis of focus measure operators for shape-from-focus. *Pattern Recognition* 46(5), 1415–1432 (2013)

[46] Abramoff, M.D., Folk, J.C., Han, D.P., *al.*: Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmology* 131(3), 351–357 (2013) DOI: 10.1001/jamaophthalmol.2013.1743