Journal of Artificial Intelligence in Medical Applications (JAIMA)
Vol. 1, No. 1, 2025, pp. 64–75
ISSN: 3106-1494 | DOI: https://doi.org/10.65511/jaima.v1i1.733

Automated Multi-Class Diabetic Retinopathy Classification Using EfficientNet-B2 Based on Color Fundus Images

Abdulrahman Abbas Mukhlif[1,*], Oana Geman[2], Omar Al-Boridi[3]
[1] Registration and Students Affairs, University Headquarter, University of Anbar, 31001, Ramadi, Anbar, Iraq; abdulrahman@uoanbar.edu.iq
[2] Data Science and AI Group, Computer Science Engineering, Chalmers and Gothenburg University, Chalmers, Sweden; geman@chalmers.se
[3] School of Engineering - RMIT University, Melbourne, Australia; omar.alboridi@rmit.edu.au

Received: 13/08/2025, Revised: 12/10/2025, Accepted: 15/10/2025, Published: 30/12/2025

ABSTRACT: Diabetic retinopathy (DR) is a leading cause of vision loss worldwide, and early detection is crucial to prevent irreversible blindness. This study aims to develop an effective deep learning framework for the automated multi-class classification of DR severity using color fundus images. We adopt EfficientNet-B2 as the backbone architecture, pre-trained on ImageNet and fine-tuned under various configurations to identify the optimal number of trainable layers. A customized preprocessing pipeline was introduced, incorporating median filtering, contrast enhancement using CLAHE, and circular cropping to emphasize retinal features while removing irrelevant background. To address class imbalance, aggressive data augmentation techniques were applied, resulting in a balanced dataset of 26,000 images across five DR stages. The model was trained and evaluated on a stratified 70:10:20 split of training, validation, and test sets. Among the fine-tuning strategies tested, unfreezing the last 100 layers of EfficientNet-B2 combined with the proposed preprocessing achieved the best performance, with 98.23% accuracy, 98.24% precision, 98.23% recall, and a 98.23% F1-score. These results demonstrate that targeted data preprocessing and selective fine-tuning significantly enhance classification accuracy in DR screening. The results confirm that combining tailored preprocessing with fine-tuned EfficientNet-B2 enables accurate and efficient multi-class DR classification. This framework offers strong potential for reliable, scalable deployment in clinical screening systems.

Keywords: Diabetic Retinopathy, Deep Learning, EfficientNet-B2, Transfer Learning, Multi-Class Classification, Fine-tuning.

1. INTRODUCTION

Diabetic Retinopathy is a progressive microvascular complication of diabetes that damages the retinal blood vessels and is a leading cause of vision impairment and blindness worldwide [1], [2], [3] . According to recent epidemiological studies, approximately 35% of individuals with diabetes exhibit signs of DR, with around 10% developing sight-threatening stages if the condition is not detected and managed promptly [4], [5]. The early stages of DR often remain asymptomatic, which underscores the importance of regular screening to enable timely therapeutic intervention .DR progresses through a series of well-defined stages, each characterized by increasing severity of retinal damage as shown in **Figure 1**. The condition is typically categorized into five classes based on the extent of microvascular abnormalities observed in fundus images (No DR, Mild, Moderate, Severe, Proliferative DR) [6].

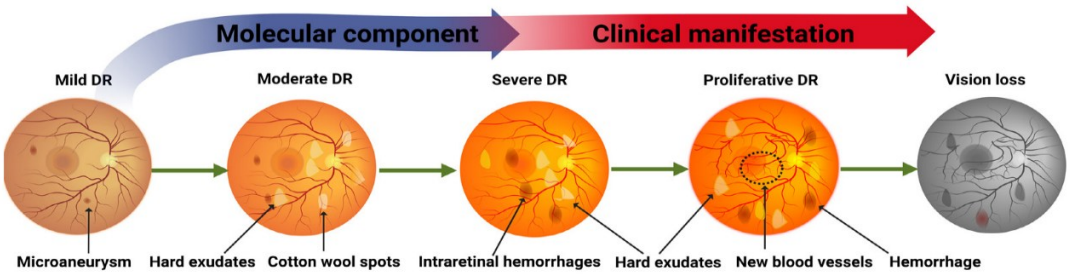


Figure 1. Illustrates how DR affects the retinal blood vessels over time, leading to vision impairment [7].

These classification labels are widely adopted in clinical practice and public datasets such as APTOS and Kaggle EyePACS [8], providing the foundation for supervised training of deep learning models aimed at automated DR severity grading. A precise recognition of these stages is essential to make an early

diagnosis and to guide treatment planning. Traditional diagnostic methods rely on manual grading of retinal fundus images by expert ophthalmologists—a labor-intensive process that is both time-consuming and susceptible to inter-observer variability [9]. This has driven significant interest in computer-aided diagnosis systems, which aim to enhance diagnostic consistency, reduce healthcare costs, and increase accessibility, especially in resource-constrained regions [10]. Deep learning, and in particular Convolutional Neural Networks, have demonstrated superior performance in medical image analysis tasks, including DR detection and classification. Comprehensive reviews confirm that CNN-based methods outperform traditional machine learning algorithms in accuracy, sensitivity, and specificity [9], [10].

Traditional methods for detecting diabetic retinopathy often rely on manual examination by ophthalmologists or basic image processing techniques, which can be time-consuming, subjective, and prone to inconsistency. These approaches may struggle to identify subtle features or early signs of DR, especially in large-scale screening scenarios. Additionally, traditional machine learning models typically require handcrafted features and lack the flexibility to adapt to the variability in fundus images, such as differences in illumination, noise, or anatomical variation. As a result, their performance in real-world clinical settings is often limited in terms of accuracy, scalability, and efficiency. Manual diagnosis of DR is often time-consuming and susceptible to human error, particularly when dealing with large-scale screening scenarios. Therefore, there is a pressing need for an automated, accurate, and computationally efficient method to classify DR severity levels. This study aims to evaluate the performance of the EfficientNet-B2 architecture in classifying DR images from a large and balanced dataset. The most significant contributions of this study can be summarized as follows:

- To develop a deep learning model for DR classification based on EfficientNet-B2, tailored for five-class severity detection.
- To design and implement a rigorous three stage preprocessing framework comprising median filtering for robust noise suppression, Contrast Limited Adaptive Histogram Equalization (CLAHE) for localized contrast enhancement, and precise circular cropping to isolate the retinal region of interest with the explicit goal of amplifying clinically relevant microvascular and hemorrhagic features and thereby substantially boosting the accuracy, sensitivity, and overall robustness of the subsequent deep learning based diabetic retinopathy classifier.
- To develop and execute a comprehensive, class specific augmentation protocol leveraging geometric transformations, intensity perturbations, and morphology-preserving operations to expand the original retinal dataset to 26,000 images with perfectly balanced representation across all diabetic retinopathy grades, thereby eliminating class skew, bolstering model generalization, and maximizing classification robustness and sensitivity.
- To systematically investigate how varying depths of fine-tuning impact classification performance by integrating a pre-trained EfficientNet-B2 with custom fully connected layers, evaluating the effects of unfreezing its last 100, 200, and 300 convolutional blocks on feature refinement, overfitting mitigation, and computational overhead.

This paper is organized as follows: related studies are given in Section 2, proposed methodology is described in Section 3, results and discussions are presented in Section 4, limitations of the study are discussed in Section 5, while Section 6 provides the conclusion of the paper and insights for future work.

2. RELATED WORKS

Several recent studies have explored the application of deep learning and transfer learning models for automated DR classification using retinal fundus images, particularly leveraging the APTOS2019 dataset and similar sources as shown in Table 1. Chilukoti, S.V. et al. (2022)[11] focused on transfer learning using EfficientNetB3 for DR detection and reported an accuracy of 87% and F1-score of 84% without applying any augmentation or balancing methods, indicating the potential of EfficientNet even under minimally preprocessed conditions. Kallel, F. et al. (2023)[12] proposed a transfer learning pipeline using multiple CNN architectures including InceptionV3, VGG16, and DenseNet169. Their results showed that InceptionV3 achieved the highest accuracy (96.88%), though recall values were relatively low for severe classes, and class imbalance was not sufficiently addressed. Akhtar, S. et al. (2024)[13] employed fine-tuned EfficientNetB3 and Xception models with extensive preprocessing steps (e.g., CLAHE, SMOTE) and achieved a test accuracy of 95.16% using EfficientNetB3, demonstrating the model's strength in capturing retinal abnormalities across five DR stages. Baskar, R. et al. (2024)[14] incorporated SMOTE and traditional data

augmentation to enhance DenseNet-169 and AlexNet-based classifiers. Although their model achieved moderate F1-scores, especially for the Severe and PDR classes, the classification of Mild and Moderate stages remained challenging, highlighting the need for more discriminative feature representations. A more recent study by Aftab, S. et al.(2025)[15] applied a data fusion strategy combining APTOS, IDRiD, and Messidor-2 datasets. Their ensemble model (EfficientNetB2, DenseNet121, ResNet50) reached a state-of-the-art accuracy of 96.96%, supported by robust preprocessing, SMOTE balancing, and comprehensive augmentation. However, the system's complexity and computational demand remain limiting factors for deployment. These studies collectively demonstrate the effectiveness of CNN-based architectures in DR classification. Nonetheless, they also reveal gaps in handling class imbalance, computational efficiency, and model generalizability. To the best of our knowledge, this study is among the first to explore the combined effect of progressive fine-tuning strategies and sequential preprocessing steps on the performance of EfficientNet-B2 in multi-class DR classification.

Table 1. Summary of studies that used transfer learning and fine-tuning to classify the APTOS2019 dataset

Author(s), Year	**Methodology**	**Dataset**	**Preprocessing Used**	**Sample Size**	**Performance**	**Key Remarks**
Chilukoti, S.V. et al., 2022 [11]	EfficientNetB3 transfer learning	Kaggle DR dataset	None Reported	~35,000	Acc: 87%, QWK: 0.85	Solid baseline; no augmentation or ensemble; VGG/ResNet underperformed
Kallel, F. et al., 2023[12]	VGG, InceptionV3, DenseNet169	APTOS 2019	Horizontal Flipping	3,662	Best: InceptionV3 Acc: 96.88%, F1: 86.6%	Good model comparison; class imbalance not addressed; test set partially used
Akhtar, S. et al., 2024[13]	EfficientNetB3 & Xception	APTOS 2019	CLAHE, SMOTE, Augmentation	3,662 → 15,410	Acc: 95.16%, Sens: 94.92%, Spec: 98.79%	High accuracy; robust preprocessing; misclassification in severe class; high computational cost
Baskar, R. et al., 2024 [14]	AlexNet & DenseNet-169	APTOS & DR Competition	SMOTE, Standard Augmentation	38,788 (augmented)	F1: 0.38–0.69 (class-wise)	Good on severe DR; poor on Mild/Moderate; class overlap issues
Aftab, S. et al., 2025 [15]	ensemble model (EfficientNetB2, DenseNet121, ResNet50)	APTOS 2019 + IDRiD + Messidor-2 (Fused)	(CLAHE + SMOTE)	26,180	Acc: 96.96%	High accuracy via ensemble + fusion

While recent AI-based approaches have shown promise in diabetic retinopathy detection, many still face challenges such as overfitting on imbalanced datasets, poor generalization to real-world images, and limited attention to subtle retinal features that distinguish early disease stages. Additionally, some models lack efficiency, making them unsuitable for deployment in large-scale or resource-constrained settings. To address these gaps, our proposed method combines a fine-tuned EfficientNet-B2 architecture with a robust preprocessing pipeline that enhances image quality and emphasizes clinically relevant regions. By balancing the dataset through aggressive augmentation and carefully controlling model complexity, our approach achieves high accuracy and reliability without sacrificing scalability or efficiency making it well-suited for practical clinical use.

3. THE PROPOSED METHODOLOGY

The methodology adopted in this study outlines a structured approach to designing, training, and evaluating a deep learning model for DR classification. It includes the key components of dataset preparation, preprocessing, model Architecture, training strategy, and performance assessment as shown in Figure 2. Each stage is designed to ensure that the model can effectively learn from the medical images and provide accurate classification results. The proposed framework ensures robustness, scalability, and relevance to real-world clinical applications.

 ISSN: 3106-1494 | **DOI**: https://doi.org/10.65511/jaima.v1i1.733

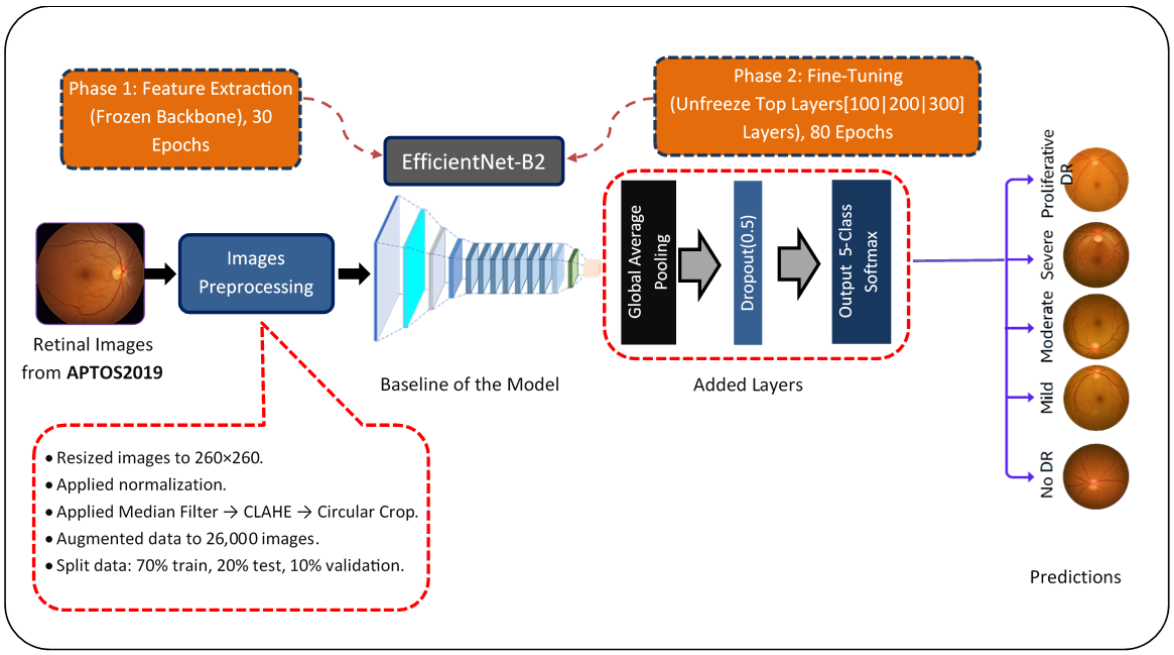


Figure 2. Steps and general framework for the proposed approach

The proposed system is a two-phase deep learning framework for multi-class classification of diabetic retinopathy severity using retinal fundus images from the APTOS2019 dataset. In the first phase, images undergo a customized preprocessing pipeline involving resizing, normalization, median filtering, CLAHE-based contrast enhancement, and circular cropping to highlight relevant retinal structures. The processed images are used to train an EfficientNet-B2 model with a frozen backbone over 30 epochs, enabling effective feature extraction. In the second phase, the model is fine-tuned by selectively unfreezing the top 100, 200, or 300 layers for an additional 80 epochs to enhance learning capacity. The architecture incorporates global average pooling, dropout (0.5), and a softmax classification layer to produce five-class predictions. To address data imbalance and improve generalization, aggressive data augmentation was applied, and the dataset was stratified into 70% training, 20% testing, and 10% validation as shown in Algorithm 1. This hybrid approach leverages both efficient preprocessing and adaptive fine-tuning to deliver high-performance retinal disease classification suitable for real-world screening environments.

Algorithm 1: Model Training Workflow (Pseudocode)

```
// Input
x ← single image (260×260×3)
// Preprocessing
function PREPROCESS(x):
    y ← RESIZE(x, 260, 260)
    y ← NORMALIZE(y)
    y ← MEDIAN_FILTER(y)
    y ← CLAHE(y)
    y ← CIRCULAR_CROP(y)
    y ← AUGMENT(y)
    return y
// Build and compile model
function BUILD_MODEL():
    base ← EfficientNetB2(no_top)
    FREEZE_LAYERS(base)
    head ← [GAP(), Dropout(0.5), Dense(softmax)]
    model ← CONNECT(base, head)
    return model
model ← BUILD_MODEL()
// Pre-training
model.PRETRAIN(train_data, val_data, epochs=30)
// Fine-tuning
for freeze_count in [100, 200, 300]:
    FREEZE_LAYERS(base, 1…freeze_count)
    UNFREEZE_LAYERS(base, freeze_count+1…end)
    model.FINETUNE(train_data, val_data, epochs=50)
end for
// Evaluation
results ← model.EVALUATE(test_data)
```

```
PRINT(results)
// Results Analysis
model.PLOT_METRICS()
// Outputs
metrics ← compute_accuracy_ Precision_ recall_F1(test_data)
confusion_heatmap ← plot_confusion(test_data)
```

End of Algorithm

3.1. DATASET

In this study, we use the APTOS 2019 Blindness Detection dataset[16] that consists of 3,662 colored retinal fundus images. The images are either classified as No DR, Mild, Moderate, Severe, or Proliferative DR to five categories of severity of DR see Table 2. There is a variability in the image collection with respect to illumination conditions, and differences in area of the retina available for detection leading to differences in image quality and contrast and in need of preprocessing.

Table 2 . APTOS 2019 Dataset Summary

Class	# Samples	Image Type	Image Resolution	Source
0 – No DR	1,805	RGB fundus photographs	Variable dimensions; heights ≈ 250–2,750 px, widths ≈ 500–4,250 px; common sizes include 2,416×1,736, 1,050×1,050, 819×614	[16]
1 – Mild DR	370			
2 – Moderate DR	999			
3 – Severe DR	193			
4 – Proliferative DR	295			
Total images:	3,662			

3.2. PREPROCESSING

To maintain data consistency and enhance the performance of the classification model, a series of pre-processing procedures were implemented in the APTOS 2019 retinal fundus images as listed in Table 3. First, all images were scaled to a standard resolution of 260x260, which was done to achieve one fixed input image resolution and so that the computational cost of applying heavy data augmentation would also parallel across all datasets. Then, pixel normalization operation was executed which scales image values to the range of [0, 1], this ensures the stability and acceleration of training deep neural networks.

The core image enhancement pipeline included three main steps:

- Median filtering was applied to reduce noise while preserving the edges of critical structures like blood vessels.
- This was followed by Contrast Limited Adaptive Histogram Equalization (CLAHE) to improve local contrast and reveal finer retinal details, especially in low-contrast regions.
- Finally, a circular cropping method was used to isolate the retinal area and remove irrelevant dark background pixels. This step ensures that the model focuses solely on the region of interest (ROI), which contains vital diagnostic features as shown in Figure 3.

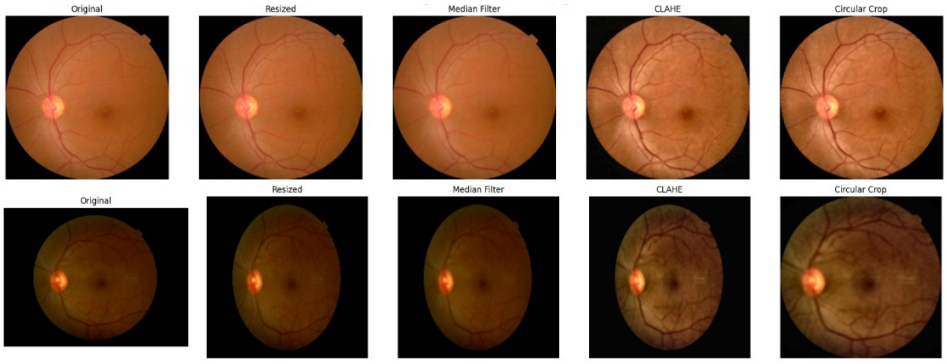


Figure 3. Retinal Image Enhancement Using Median Filter, CLAHE, and Circular Crop

To enhance generalization and prevent overfitting, data augmentation techniques were employed, including random rotation (±10 degrees), zooming (up to 20%), and horizontal flipping as shown in Figure 4. These augmentations were applied during training to synthetically increase the diversity of the dataset.

 ISSN: 3106-1494 | **DOI**: https://doi.org/10.65511/jaima.v1i1.733

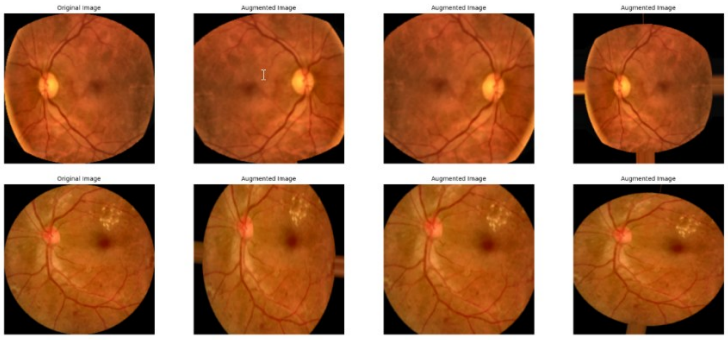


Figure 4 .Examples of Original and Augmented Retinal Fundus Images

To address class imbalance in the original dataset, all five disease categories were balanced to contain exactly 5,200 images each as shown in Figure 5, bringing the total number of samples to 26,000. This was achieved by augmenting underrepresented classes using the transformations described earlier.

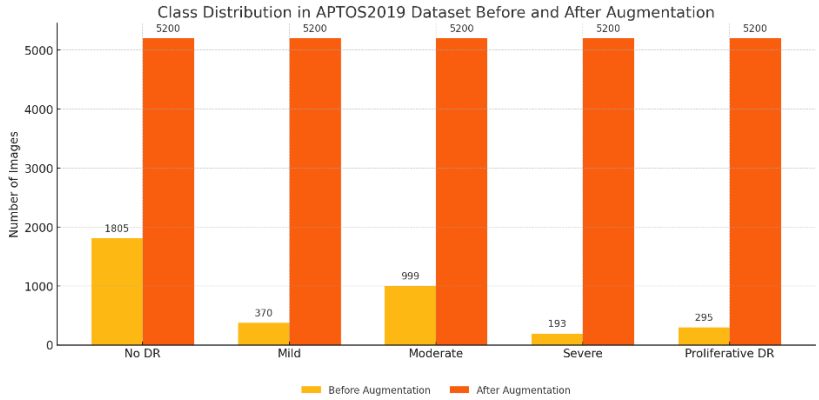


Figure 5. Class Distribution in APTOS2019 Dataset Before and After Data Augmentation

Finally, the dataset was split into 70% training, 10% validation, and 20% testing sets using stratified sampling to ensure consistent class distribution across splits.

Table 3 . Summary of Preprocessing Steps

Preprocessing Step	**Description**
Image Resizing	All images resized to 260×260 pixels
Normalization	Pixel values scaled to the range [0, 1]
Median Filter → CLAHE → Circular Crop	Noise reduction, contrast enhancement, and region of interest isolation
Data Augmentation	Rotation (10°), Zoom (0.2), Horizontal Flip (True)
Balancing	Increased to 26,000 images (5,200 per class)
Data Splitting	70% training, 20% testing, 10% validation

3.3. MODEL ARCHITECTURE

The proposed deep learning model for DR classification leverages the EfficientNet-B2 architecture as its backbone due to its ability to achieve high accuracy with reduced computational complexity. The input to the model consists of color retinal fundus images, resized to a uniform dimension of 260×260×3 pixels to match the input specifications of the backbone. The EfficientNet-B2 model was initialized with pre-trained ImageNet weights and employed in a two-phase training strategy. During the first phase, all base layers were kept frozen to preserve the learned generic visual features, and only the custom classification layers were trained. In the second phase, we performed fine-tuning by gradually unfreezing layers of the backbone network to allow the model to adapt more effectively to the domain-specific retinal data.

To identify the optimal fine-tuning configuration, we experimented with multiple unfreezing depths by selectively unfreezing the top 100, 200, and 300 layers of the EfficientNet-B2 backbone. This comparative approach allowed us to determine the most effective layer depth for extracting retinal disease-related features while maintaining training stability and avoiding overfitting. To tailor the model for the five-class

classification task, a series of custom layers were appended to the base model as shown in Table 4. These included the following:

- Global Average Pooling layer to reduce the spatial dimensions of the feature maps.
- A Dropout layer with a dropout rate of 0.5 to prevent overfitting.
- A final Output Layer with 5 units and Softmax activation to perform multi-class classification as shown in Figure 6.

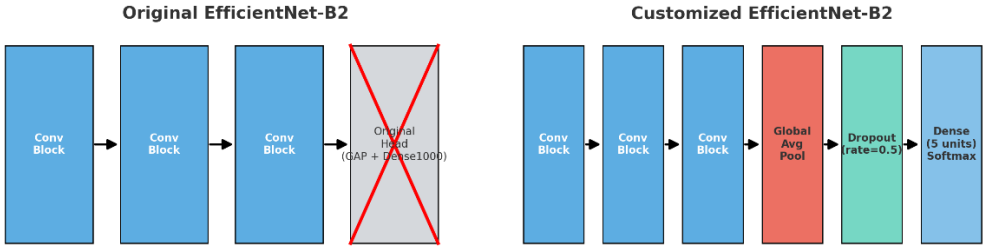


Figure 6 . Original vs. Customized EfficientNet-B2 Architecture

This architectural design ensures a progressive reduction in feature dimensionality while maintaining robust learning through normalization and regularization techniques.

Table 4 . Layers added to the EfficientNet-B2 model structure

Layer	Units	Activation	Dropout Rate	Function
Global Average Pooling	1280	-	-	Dimensionality Reduction
Dropout	-	-	0.5	Prevent overfitting
Output Layer	5	Softmax	-	Final Classification

3.4. TRAINING

The training process of the proposed model was carried out in two sequential phases as shown in Table 5. In the first phase, the EfficientNet-B2 backbone was kept frozen, and only the custom classification layers were trained. This approach leveraged the generalized pre-trained features while allowing the newly added top layers to adapt to the DR classification task. A learning rate of 1e-3 was used in this stage. In the second phase, extensive fine-tuning was conducted by unfreezing different numbers of layers in the EfficientNet-B2 backbone. Specifically, three scenarios were evaluated by unfreezing the top 100, 200, and 300 layers respectively. Each fine-tuning scenario was applied both with and without the preprocessing pipeline that includes Median Filter → CLAHE → Circular Crop, to evaluate the impact of image enhancement on feature learning and model generalization. A reduced learning rate of 1e-5 was used in this phase to perform gradual updates to the pre-trained weights. Training was performed using the Adam optimizer due to its adaptive learning rate properties. The total number of epochs was set to 80 with a batch size of 32. Additionally, EarlyStopping and ReduceLROnPlateau callbacks were incorporated to prevent overfitting and dynamically adjust the learning rate when validation performance plateaued.

Table 5. Training Configuration

Component	**Value / Configuration**
Training Phases	Phase 1: Freeze base, Phase 2: Fine-tune (unfreezing the top 100, 200, and 300 layers)
Optimizer	Adam
Learning Rate (Phase 1)	1e-3
Learning Rate (Phase 2)	1e-5
Batch Size	32
Epochs	80
Callbacks	EarlyStopping, ReduceLROnPlateau

3.5. EVALUATION

Numerous classification metrics were used to evaluate model performance so that a holistic view of its accuracy could be obtained. These standard metrics are accuracy, precision, recall, F1-score that are widely used in the context of multi-class classification problem [17], [18], [19]. Each of these metrics are calculated as:

 ISSN: 3106-1494 | **DOI**: https://doi.org/10.65511/jaima.v1i1.733

Accuracy: is the ratio of correctly classified examples to all the examples in the dataset:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Precision: It is defined as the ratio of the number of true positive observations, to the total number of positives returned by the classifier.

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

Recall (Sensitivity): the number of true positive observations divided by all positive observations:

$$\text{Recall} = \frac{TP}{(TP + FN)}$$

F1-Score: is the harmonic mean of precision & recall, it gives a balance b/w the both of them:

$$\text{F1-Score} = \frac{2 * (\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

Where:
- TP: True Positives
- TN: True Negatives
- FP: False Positives
- FN: False Negatives

Moreover, the confusion matrix was examined for assessing the ability of the model to differentiate between the five DR classes. This matrix gives us an idea about how much of the errors are influencing the result of our model and what are the good and bad class-wise performance.

4. RESULTS AND DISCUSSION

Table 6 summarizes the classification results of different training techniques and preprocessing situations existing in the EfficientNet-B2 model with comparisons for the APTOS 2019. The baseline model was the EfficientNet-B2 with all base layers frozen and no preprocessing and achieved an accuracy of 71.90%. Using preprocessing (Median Filter → CLAHE → Circular Crop) without tuning slightly decreased description performance (70.44%) probably because the model did not fit well to the new image distribution with frozen weights.

Table 6 . Performance metrics for each training and preprocessing scenario using EfficientNetB2

Methods applied to the EfficientNet-B2 model	Accuracy	Precision	Recall	F1-score
freeze all the base-model layers	71.90%	71.60%	71.90%	71.65%
freeze all the base-model layers + Median Filter → CLAHE → Circular Crop	70.44%	70.75%	70.44%	70.14%
Fine-Tuning(unfreeze the last 300 layers)	97.23%	97.26%	97.23%	97.22%
Fine-Tuning(unfreeze the last 300 layers) + Median Filter → CLAHE → Circular Crop	97.69%	97.71%	97.69%	97.69%
Fine-Tuning(unfreeze the last 200 layers)	98.08%	98.09%	98.08%	98.08%
Fine-Tuning(unfreeze the last 200 layers) + Median Filter → CLAHE → Circular Crop	98.12%	98.13%	98.12%	98.11%
Fine-Tuning(unfreeze the last 100 layers)	97.94%	97.95%	97.94%	97.94%
Fine-Tuning(unfreeze the last 100 layers) + Median Filter → CLAHE → Circular Crop	**98.23%**	**98.24%**	**98.23%**	**98.23%**

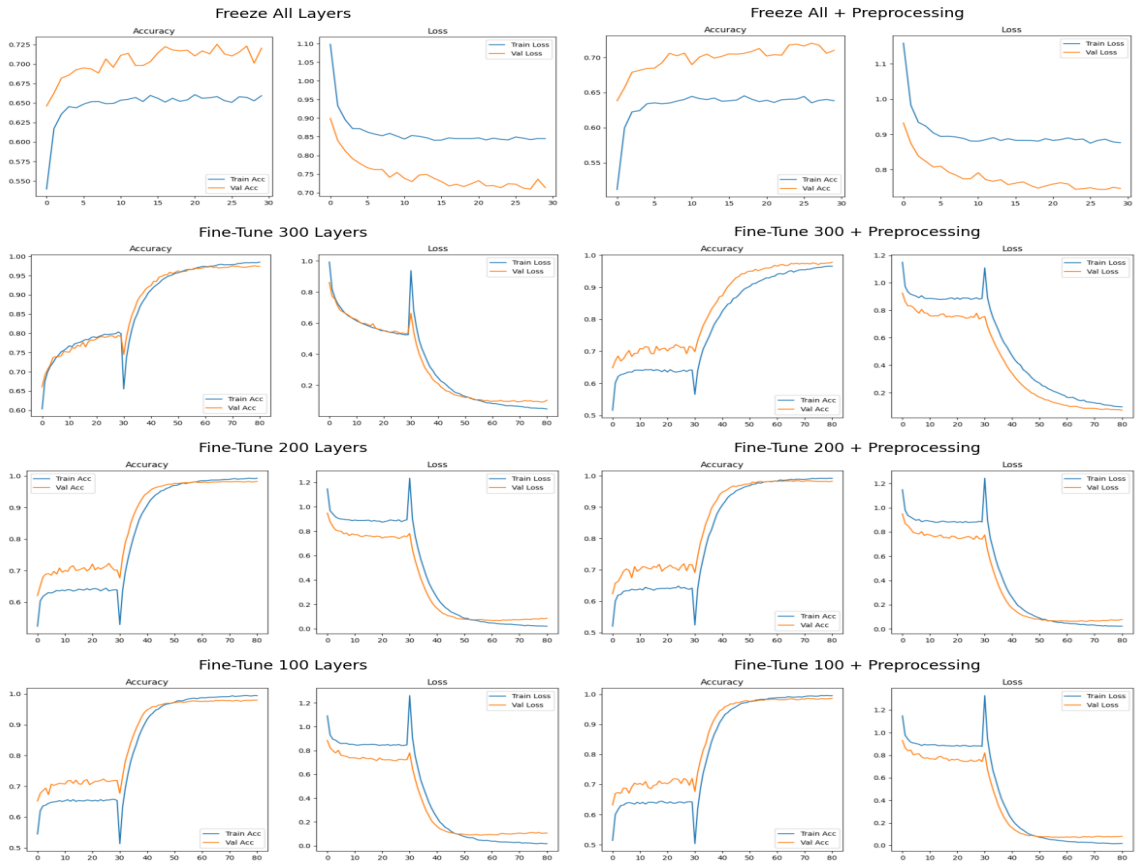
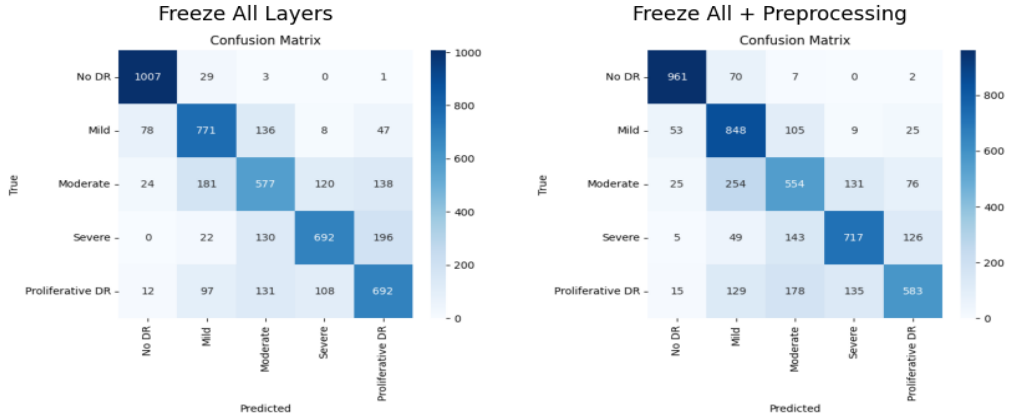


Figure 7 . Training and Validation Accuracy and Loss Curves for Each Fine-Tuning Strategy

A significant improvement was observed with fine-tuning. Unfreezing the last 300 layers led to an accuracy of 97.23%, which improved further to 97.69% when combined with preprocessing. Similarly, fine-tuning the last 200 layers yielded 98.08% accuracy without preprocessing and 98.12% with preprocessing. The best results were achieved when fine tuning only the last 100 layers, reaching an accuracy of 98.23% with preprocessing, along with precision, recall, and F1-score all exceeding 98%. These results suggest that:

- Fine-tuning even fewer layers (100) can outperform deeper fine-tuning (200 or 300 layers), likely due to better generalization and fewer trainable parameters.
- The proposed preprocessing pipeline (Median Filter → CLAHE → Circular Crop) enhances performance across all fine-tuning scenarios.
- The combination of light fine-tuning and robust preprocessing yields the most accurate model, confirming the effectiveness of our methodology as shown in Figure 8.



 ISSN: 3106-1494 | **DOI**: https://doi.org/10.65511/jaima.v1i1.733

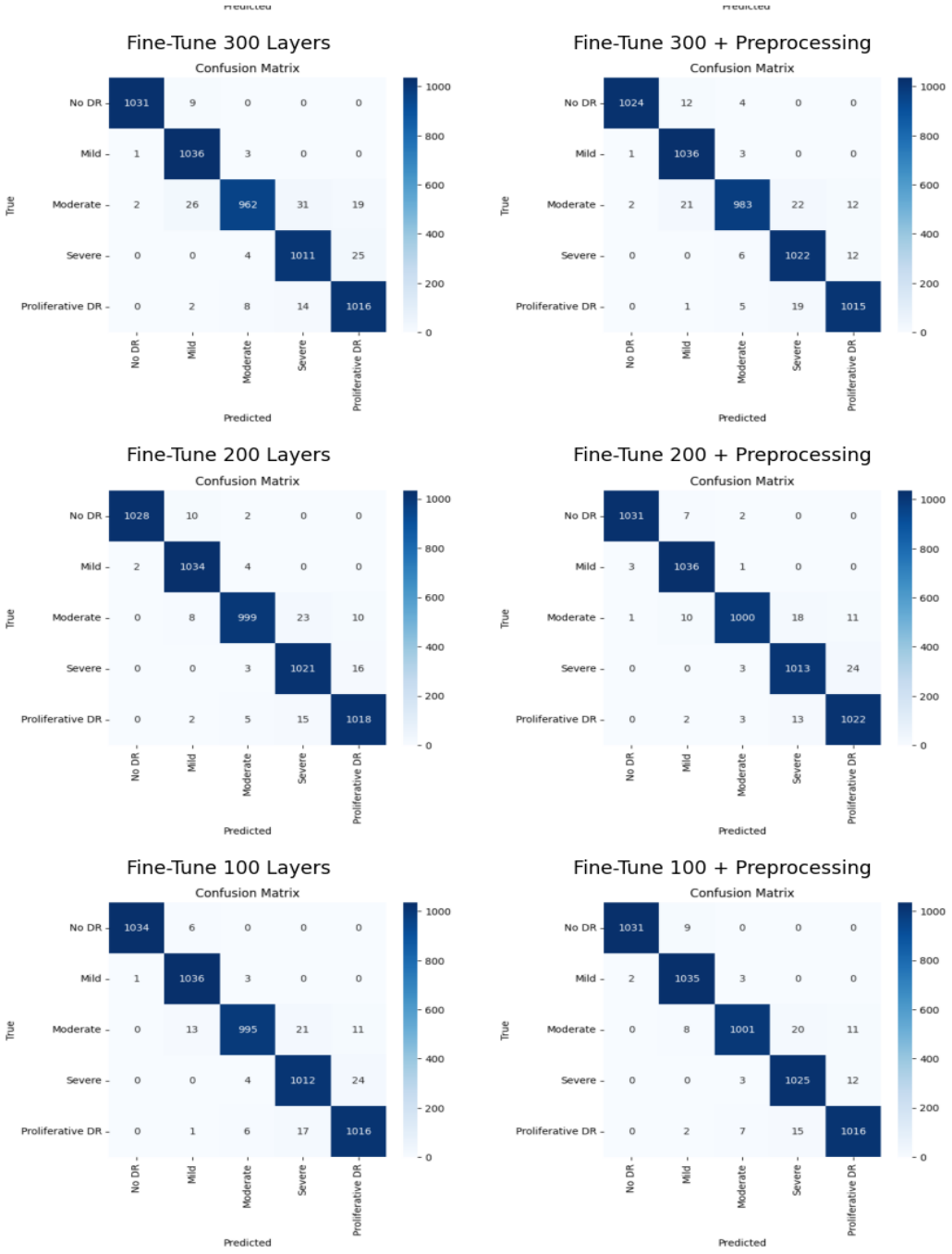


Figure 8 .Confusion Matrices for Different Fine-Tuning Scenarios on EfficientNet-B2

To contextualize our findings, we compared them with recent studies on the same or similar datasets such as Chilukoti et al. (2022) [11] employed EfficientNetB3 on the Kaggle DR Dataset and achieved 87% accuracy with a QWK of 0.85. However, their study lacked preprocessing and augmentation, which limited model generalization. In contrast, our model integrates a comprehensive preprocessing pipeline and image balancing strategy, leading to a significantly higher accuracy of 98.23%. Kallel et al. (2023) [12] reported 96.88% accuracy using InceptionV3 on APTOS 2019, applying only horizontal flipping. While their results are competitive, their preprocessing was minimal and class imbalance remained unaddressed. Our study outperforms this by implementing targeted preprocessing and balancing all classes to 5,200 samples each, eliminating bias and improving performance consistency across underrepresented classes. Akhtar et al. (2024) [13] utilized EfficientNetB3 and Xception with an advanced pipeline including CLAHE, SMOTE, and heavy augmentation. Their model achieved 95.16% accuracy and excellent sensitivity/specificity. However,

the method suffered from misclassifications in the severe class and imposed a high computational burden due to model complexity. Our approach achieves higher accuracy with more efficient computation using only EfficientNetB2, while still addressing class confusion through focused fine-tuning and balanced training data.

Also, Baskar et al. (2024) [14] implemented AlexNet and DenseNet-169 on a combined dataset, but the F1-scores varied significantly by class (0.38–0.69), particularly performing poorly on Mild and Moderate classes. The performance gaps were attributed to class overlaps and imbalance. In contrast, our method yields stable, high F1-scores across all classes by leveraging class balancing and preprocessing that emphasizes lesion contrast. Aftab et al. (2025) [15] proposed an ensemble of EfficientNetB2, DenseNet121, and ResNet50, trained on a fused dataset (APTOS, IDRiD, Messidor-2) and achieved 96.96% accuracy. Although impressive, their model's reliance on ensemble complexity and multiple datasets increases the reproducibility barrier. Our single-model approach matches and exceeds this performance (98.23%) on APTOS 2019 alone as shown in Table 7, showcasing that strategic preprocessing and scenario-based fine-tuning can yield comparable, if not superior, results without fusion or ensembling.

Table 7 .Performance Comparison with Related Works

Researcher(s)	Year	Accuracy	Precision	Recall	F1-Score
Chilukoti, S.V. et al. [11]	2022	87.00%	85.00%	87.00%	84.00%
Kallel, F. et al. [12]	2023	96.88%	92.00%	85.40%	86.60%
Akhtar, S. et al.[13]	2024	95.16%	-	94.92%	95.24%
Baskar, R. et al. [14]	2024	-	-	-	38%-69%
Aftab, S. et al. [15]	2025	96.96%	-	96.93%	96.95%
Proposed work	2025	**98.23%**	**98.24%**	**98.23%**	**98.23%**

Its clear that our study contributes a comprehensive yet computationally efficient framework for DR classification that rivals or exceeds the performance of more complex or resource-intensive methods. The effectiveness stems not from the architecture alone, but from a synergistic combination of preprocessing, class balancing, and staged fine-tuning—a strategy validated through systematic experimentation.

5. LIMITATIONS

This study has a few limitations. First, the model was evaluated only on the APTOS 2019 dataset, which may limit its generalizability to other datasets or clinical environments. Second, the preprocessing pipeline, while effective, adds computational complexity. Third, no external validation was performed, which restricts the assessment of real-world performance. Lastly, although multiple fine-tuning scenarios were explored, further optimization may still enhance results.

6. CONCLUSION

To maintain data consistency and enhance the performance of the classification model, a series of pre-processing procedures were implemented in the APTOS 2019 retinal fundus images as listed in Table 3 This study presents a deep learning approach for DR classification based on the EfficientNet-B2 architecture. The proposed method integrates a custom preprocessing pipeline involving median filtering, CLAHE, and circular cropping, in addition to data augmentation techniques to balance the dataset by increasing each class to 5,200 images, resulting in a total of 26,000 images. Multiple fine-tuning strategies were explored by unfreezing the last 100, 200, and 300 layers of the pre-trained model. The optimal performance was achieved when unfreezing the last 100 layers combined with the preprocessing pipeline. This configuration resulted in a remarkable performance with 98.23% accuracy, 98.24% precision, 98.23% recall, and 98.23% F1-score, demonstrating the effectiveness of the proposed strategy in multi-class classification tasks. For future works it's better to use external validation & domain generalization via assess additional external retinal datasets like IDRiD, Messidor-2, and OCTA scans to train/consolidate domain-adaptation techniques and adversarial feature alignment to reduce dataset shift. Also, using Explainable AI (XAI) by implementing Grad-CAM techniques and SHAP or LIME to explain AI behavior by attributing pixels or features aid to deeper model layers. Even more, EXI-AI which enables clinicians to control guided model probing for explanation can be researched. Implementing these steps will significantly enhance the accuracy, efficiency, and interpretability of our screening system for real-world applications in diabetic retinopathy.

FUNDING INFORMATION

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

COFLICTS OF INTERESTS

The authors declare that there is no conflict of interest regarding the publication of this paper.

DATA AVAILABILITY STATEMENTS

The dataset used in this study is publicly available at the following link: https://kaggle.com/competitions/aptos2019-blindness-detection

REFERENCES

[1] M. W. Nadeem, H. G. Goh, M. Hussain, S. Y. Liew, I. Andonovic, and M. A. Khan, "Deep Learning for Diabetic Retinopathy Analysis: A Review, Research Challenges, and Future Directions," *Sensors (Basel),* vol. 22, no. 18, Sep 8 2022, doi: 10.3390/s22186780.

[2] T. E. Tan and T. Y. Wong, "Diabetic retinopathy: Looking forward to 2030," *Front Endocrinol (Lausanne),* vol. 13, p. 1077669, 2022, doi: 10.3389/fendo.2022.1077669.

[3] A. Senapati, H. K. Tripathy, V. Sharma, and A. H. Gandomi, "Artificial intelligence for diabetic retinopathy detection: A systematic review," *Informatics in Medicine Unlocked,* vol. 45, 2024, doi: 10.1016/j.imu.2024.101445.

[4] C. P. Cheyne *et al.*, "Incidence of sight-threatening diabetic retinopathy in an established urban screening programme: An 11-year cohort study," *Diabet Med,* vol. 38, no. 9, p. e14583, Sep 2021, doi: 10.1111/dme.14583.

[5] Z. L. Teo *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.

[6] W. L. Alyoubi, M. F. Abulkhair, and W. M. Shalash, "Diabetic Retinopathy Fundus Image Classification and Lesions Localization System Using Deep Learning," *Sensors (Basel),* vol. 21, no. 11, May 26 2021, doi: 10.3390/s21113704.

[7] V. Kaushik, L. Gessa, N. Kumar, and H. Fernandes, "Towards a New Biomarker for Diabetic Retinopathy: Exploring RBP3 Structure and Retinoids Binding for Functional Imaging of Eyes In Vivo," *Int J Mol Sci,* vol. 24, no. 5, Feb 23 2023, doi: 10.3390/ijms24054408.

[8] A. A. Mukhlif, B. Al-Khateeb, and M. A. Mohammed, "An extensive review of state-of-the-art transfer learning techniques used in medical imaging: Open issues and challenges," *Journal of Intelligent Systems,* vol. 31, no. 1, pp. 1085-1111, 2022, doi: 10.1515/jisys-2022-0198.

[9] A. Skouta, A. Elmoufidi, S. Jai-Andaloussi, and O. Ouchetto, "Deep learning for diabetic retinopathy assessments: a literature review," *Multimedia Tools and Applications,* vol. 82, no. 27, pp. 41701-41766, 2023/11/01 2023, doi: 10.1007/s11042-023-15110-9.

[10] N. Asiri, M. Hussain, F. Al Adel, and N. Alzaidi, "Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey," *Artificial Intelligence in Medicine,* vol. 99, p. 101701, 2019/08/01/ 2019, doi: https://doi.org/10.1016/j.artmed.2019.07.009.

[11] S. V. Chilukoti, A. Maida, and X. Hei, *Diabetic Retinopathy Detection using Transfer Learning from Pre-trained Convolutional Neural Network Models*. 2022.

[12] F. Kallel and A. Echtioui, "Retinal fundus image classification for diabetic retinopathy using transfer learning technique," *Signal, Image and Video Processing,* vol. 18, no. 2, pp. 1143-1153, 2023, doi: 10.1007/s11760-023-02820-8.

[13] S. Akhtar, S. Aftab, M. Ahmad, and A. Akhtar, "Diabetic Retinopathy Severity Grading UsingTransfer Learning Techniques," *International Journal of Engineering and Manufacturing,* vol. 14, no. 6, pp. 41-53, 2024, doi: 10.5815/ijem.2024.06.04.

[14] R. Baskar, E. Sabu, and C. Mazo, "Deep CNNs for Diabetic Retinopathy Classification: A Transfer Learning Perspective," presented at the 2024 IEEE International Symposium on Biomedical Imaging (ISBI), 2024.

[15] S. Aftab and S. Akhtar, "Diabetic Retinopathy Severity Classification Using Data Fusion and Ensemble Transfer Learning," *Journal of Software Engineering and Applications,* vol. 18, no. 01, pp. 1-23, 2025, doi: 10.4236/jsea.2025.181001.

[16] Karthik, Maggie, and S. Dane. "APTOS 2019 Blindness Detection." Kaggle. https://kaggle.com/competitions/aptos2019-blindness-detection (accessed.

[17] A. A. Nafea, M. S. Ibrahim, A. A. Mukhlif, M. M. Al-Ani, and N. Omar, "An Ensemble Model for Detection of Adverse Drug Reactions," *Aro-the Scientific Journal of Koya University,* vol. 12, no. 1, pp. 41-47, 2024, doi: 10.14500/aro.11403.

[18] S. Al-Fahdawi *et al.*, "Fundus-DeepNet: Multi-label deep learning classification system for enhanced detection of multiple ocular diseases through data fusion of fundus images," *Information Fusion,* vol. 102, 2024, doi: 10.1016/j.inffus.2023.102059.

[19] S. Lal *et al.*, "Adversarial Attack and Defence through Adversarial Training and Feature Fusion for Diabetic Retinopathy Recognition," *Sensors (Basel),* vol. 21, no. 11, Jun 7 2021, doi: 10.3390/s21113922.