Optimizing Diabetic Retinopathy Classification Using EfficientNet-B3 with Data Augmentation and Oversampling

A A JE Veggy Priyangka[1*], Tuga Mauritsius[2]

[1,2]*Information Systems Management Department, Binus Graduate Program, Bina Nusantara University, Jakarta, Indonesia*

*corr-author: a.priyangka@binus.ac.id

Abstract - Diabetic retinopathy (DR) is a leading cause of preventable blindness among diabetes patients. This study optimizes DR severity classification using EfficientNet-B3 with transfer learning combined with data handling strategies. Using the APTOS 2019 dataset containing 3,662 retinal fundus images across five severity classes, three experimental scenarios were evaluated: (1) baseline CNN, (2) CNN with data augmentation, and (3) CNN with data augmentation and random oversampling. Performance was measured using Quadratic Weighted Kappa (QWK), accuracy, precision, recall, F1-score, and ROC-AUC. Results demonstrate that Scenario III achieves the best performance with QWK of 0.8496 and accuracy of 77.00%, representing significant improvement over baseline (QWK: 0.4998) and augmentation-only models (QWK: 0.5728). The combination of data augmentation and random oversampling effectively addresses class imbalance in medical image datasets. This study provides empirical evidence on combining transfer learning with data balancing strategies for automated DR screening systems.

Keywords: Diabetic retinopathy, deep learning, EfficientNet-B3, data augmentation, random oversampling.

I. INTRODUCTION

Diabetes mellitus (DM) has become a global health crisis with an estimated 537 million adults living with diabetes worldwide in 2021, projected to reach 783 million by 2045 [1]. Indonesia ranks fifth globally with approximately 19.5 million diabetes cases, representing a significant public health burden [2]. One of the most serious complications of diabetes is diabetic retinopathy (DR), a microvascular complication affecting the retina that can lead to irreversible vision loss if left untreated [3].

Diabetic retinopathy develops progressively through distinct stages, from mild non-proliferative changes to severe proliferative disease requiring immediate intervention [4]. The International Clinical Diabetic Retinopathy Disease Severity Scale classifies DR into five levels: No DR, Mild NPDR, Moderate NPDR, Severe NPDR, and Proliferative DR [5]. Early detection at mild to moderate stages enables timely treatment and significantly reduces the risk of vision loss [6].

The gold standard for DR diagnosis involves fundus photography examination by trained ophthalmologists [7]. However, this approach faces significant challenges in resource-limited settings. Developing countries, including Indonesia, face a severe shortage of eye care professionals relative to their large populations [8]. This disparity highlights the urgent need for automated screening systems to support early detection efforts.

Deep learning, particularly Convolutional Neural Networks (CNNs), has demonstrated remarkable success in medical image analysis tasks [9]. Transfer learning approaches, where models pre-trained on large datasets like ImageNet are fine-tuned for specific medical tasks, have shown superior performance compared to training from scratch [10]. Among various CNN architectures, EfficientNet has gained attention for its balanced scaling approach, achieving state-of-the-art accuracy with fewer parameters [11]. Recent studies have demonstrated the effectiveness of EfficientNet variants for diabetic retinopathy severity classification [12]. Furthermore, approaches addressing class imbalance in retinal image analysis have shown promising results with improved detection of minority classes [13].

However, medical image datasets commonly suffer from class imbalance, where certain disease classes are underrepresented [14]. The APTOS 2019 dataset exhibits significant imbalance with No DR class comprising approximately 49% of samples while Mild DR represents only 7% [15]. This imbalance can bias models toward majority classes, reducing sensitivity for detecting minority severity levels [16].

Several studies have addressed DR severity classification using deep learning. Gulshan et al. [17] demonstrated the feasibility of deep learning for DR

screening with high sensitivity and specificity. Subsequent studies explored various architectures including VGGNet [18] and ResNet [19] for this task. Recent work has achieved competitive results using EfficientNet architectures [20]. However, comprehensive evaluation of data handling strategies for imbalanced DR datasets remains limited.

Data augmentation and oversampling represent two complementary approaches for addressing limited and imbalanced data [21]. Data augmentation generates synthetic variations of existing samples through transformations such as rotation, flipping, and color adjustments [22]. Oversampling techniques duplicate or synthesize minority class samples to balance class distributions [23]. While SMOTE (Synthetic Minority Over-sampling Technique) is widely used for tabular data, random oversampling remains effective for image classification where feature generation is handled by deep networks [24].

This study addresses the following research question: How do data augmentation and random oversampling strategies affect the performance of EfficientNet-B3 for diabetic retinopathy severity classification? The objectives are: (1) to evaluate the baseline performance of EfficientNet-B3 with transfer learning on DR classification, (2) to assess the impact of data augmentation on model generalization, and (3) to determine the effectiveness of combining data augmentation with random oversampling for handling class imbalance.

The contributions of this research include: (1) systematic empirical investigation demonstrating that random oversampling combined with data augmentation achieves superior QWK improvement (70% over baseline) compared to augmentation alone (14.6%), providing quantitative evidence for practitioners selecting data handling strategies; (2) novel finding that the accuracy-QWK trade-off is acceptable for ordinal medical image classification—Scenario III achieves lower accuracy (77%) but substantially higher QWK (0.8496), challenging the conventional focus on accuracy metrics; (3) practical validation of a computationally efficient approach using CPU-only training, demonstrating that effective DR screening models can be developed without expensive GPU infrastructure, which is particularly relevant for resource-constrained healthcare settings in developing countries.

These findings are significant for healthcare informatics, particularly in resource-constrained settings where automated screening can supplement limited specialist availability.

II. METHOD

A. Research Design

This study employs an experimental research design to systematically evaluate the impact of data handling strategies on deep learning performance for diabetic retinopathy classification. The independent variables in this study include: (1) data augmentation strategy (applied/not applied), and (2) random oversampling technique (applied/not applied). These two data handling strategies are systematically varied to assess their individual and combined effects on model performance. The dependent variables are classification performance metrics consisting of Quadratic Weighted Kappa (QWK) as the primary metric, along with accuracy, precision, recall, F1-score, and ROC-AUC as secondary metrics.

To systematically evaluate the impact of each data handling strategy, three experimental scenarios are designed as follows as shown in Table I.

Scenario I (Baseline) serves as the control condition where EfficientNet-B3 with transfer learning from ImageNet is trained directly on the original imbalanced dataset without any data handling intervention. This scenario establishes the baseline performance against which improvements from data handling strategies are measured.

TABLE I
EXPERIMENTAL SCENARIO CONFIGURATIONS

Scenario	Data Augmentation	Random Oversampling	Description
I	Not Applied	Not Applied	Baseline model using EfficientNet-B3 with transfer learning only, without any data handling strategies
II	Applied	Not Applied	EfficientNet-B3 with real-time data augmentation to improve generalization
III	Applied	Applied	EfficientNet-B3 with both data augmentation and random oversampling to address class imbalance

Scenario II (Data Augmentation Only) introduces real-time data augmentation during training to artificially expand the training set through geometric and photometric transformations. This scenario evaluates the effect of augmentation on model generalization without addressing the class imbalance problem.

Scenario III (Data Augmentation + Random Oversampling) combines both strategies by first applying random oversampling to balance class distributions, followed by real-time augmentation during training. This scenario tests the hypothesis that addressing class imbalance complements augmentation benefits for improved classification performance.

This section outlines the process stages undertaken in this research, including data preparation, preprocessing, data handling strategies, model development, and performance evaluation. The flowchart of the diabetic retinopathy classification process in this study is presented in Fig. 1.

The research workflow follows the CRISP-DM methodology [25] phases as illustrated in Fig. 1. The Data Understanding phase involves loading the APTOS 2019 dataset and analyzing class distributions to identify the imbalance problem (Section B). The Data Preparation phase encompasses three sequential processes: (1) preprocessing, where raw images are resized to 224×224 pixels and normalized (Section C); (2) data augmentation, where geometric and photometric transformations are applied to increase training diversity (Section D); and (3) random oversampling, where minority class samples are duplicated to balance class distributions (Section E). The Modeling phase involves constructing the EfficientNet-B3 architecture with transfer learning and training using the two-phase strategy (Sections F and G). Finally, the Evaluation phase computes performance metrics including QWK, accuracy, precision, recall, F1-score, and ROC-AUC on the validation set (Section H). The workflow transitions proceed from raw images as input, through preprocessed images, to augmented and oversampled training set, then to trained model, followed by predictions, and finally to performance metrics as output. Each scenario (I, II, III) follows the same workflow but with different combinations of data handling strategies enabled, allowing systematic comparison of their effects.

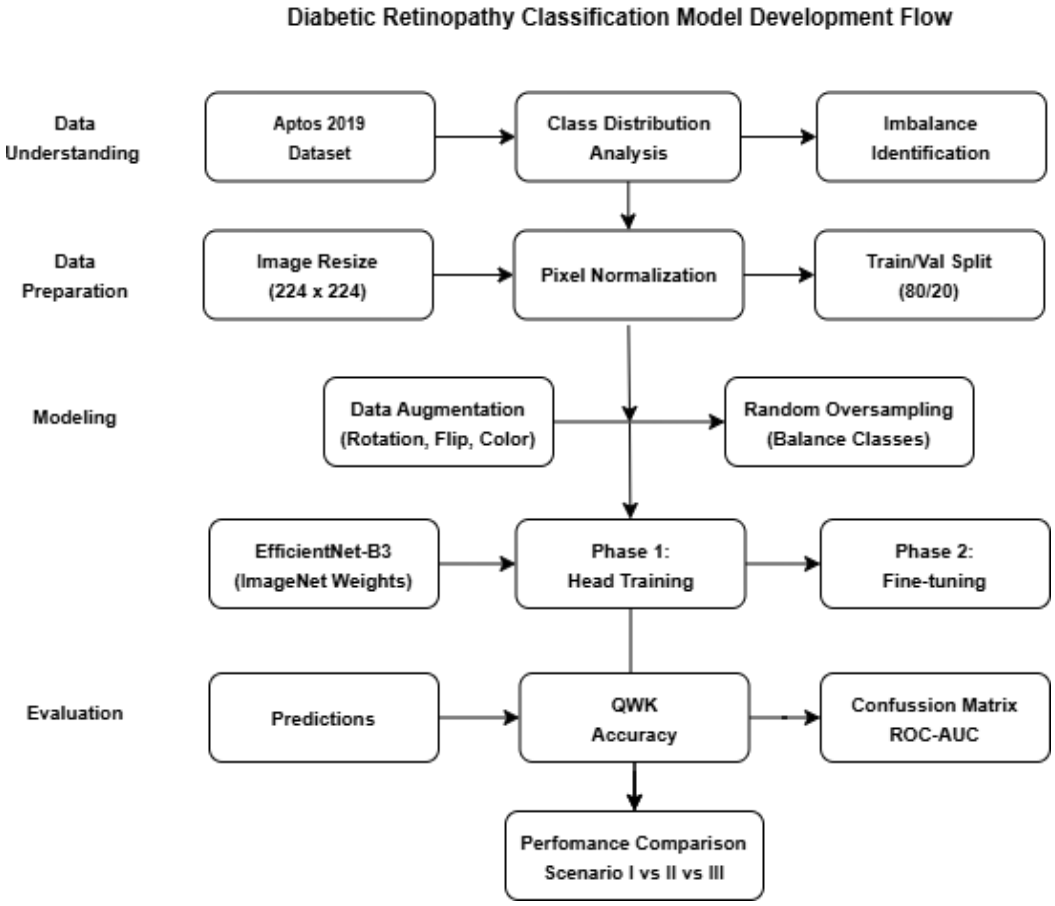

Fig. 1 Diabetic retinopathy classification model development flow

B. Dataset

The APTOS 2019 Blindness Detection dataset from Kaggle [26] is used in this study. The dataset contains 3,662 high-resolution retinal fundus images captured under various clinical conditions. Images are labeled by clinical experts according to the International Clinical DR Severity Scale into five classes, as shown in Table II.

The dataset exhibits significant class imbalance with imbalance ratio (majority/minority) of 9.35:1 between No DR and Severe NPDR classes. This imbalance necessitates appropriate data handling strategies to prevent model bias. As shown in Table II, the No DR class dominates nearly half of the dataset (49.29%), while the Severe NPDR class represents the smallest proportion with only 5.27% of total samples. The Moderate NPDR class constitutes the second largest group at 27.28%, followed by Mild NPDR (10.10%) and Proliferative DR (8.06%). This distribution pattern reflects real-world clinical scenarios where severe cases are relatively rare compared to healthy or mild conditions, presenting a common challenge in medical image classification tasks.

The dataset was split using stratified sampling with an 80/20 ratio for training and validation sets respectively, ensuring proportional representation of all severity classes in both subsets. A separate test set was not created as the validation set serves dual purposes for hyperparameter tuning and final evaluation, following common practice in Kaggle competition datasets where external test labels are unavailable [26].

C. Data Preprocessing

The original APTOS 2019 dataset contains retinal fundus images with varying resolutions, ranging from approximately 474×358 to 3388×2588 pixels. To match the input requirements of EfficientNet-B3 architecture, all images undergo the following preprocessing steps.

All images are resized to 224×224 pixels using bilinear interpolation. This target size was selected to match EfficientNet-B3's standard input dimension while balancing between preserving diagnostic features and computational efficiency.

Direct resizing without explicit aspect ratio preservation is applied. While this may introduce minor geometric distortion, retinal fundus images in the APTOS 2019 dataset are approximately circular and centered within the frame, minimizing the impact on clinically relevant features such as microaneurysms, hemorrhages, and exudates.

No explicit retinal region cropping was performed prior to resizing, as the APTOS 2019 images are already pre-centered on the retinal area with minimal peripheral artifacts. The circular retinal region occupies the majority of each image frame.

Pixel values are normalized from the original range [0, 255] to [0, 1] by dividing by 255 as in (1).

$$x_{\text{rescaled}} = \frac{x_{\text{original}}}{255} \quad (1)$$

where $x_{original}$ represents the original pixel intensity value in the range [0, 255], and $x_{rescaled}$ is the normalized value in the range [0, 1].

All images are ensured to be in RGB format (3 channels) to match the model's input requirements. Grayscale or RGBA images are converted accordingly.

This preprocessing pipeline ensures consistent input dimensions and value ranges suitable for neural network training while preserving the diagnostic information present in the original fundus images.

D. Data Augmentation

For Scenarios II and III, real-time data augmentation is applied during training to increase training diversity and improve model generalization. The augmentation techniques applied in this study are summarized in Table III.

E. Random Oversampling

For Scenario III, random oversampling is applied to balance class distributions before training. Random oversampling is applied to the training set only (after 80/20 train-validation split).

The oversampling factor k_c for each class c is calculated as (2).

$$k_c = \left\lceil \frac{N_{max}^{train}}{N_c^{train}} \right\rceil \quad (2)$$

TABLE II
DISTRIBUTION OF DR SEVERITY CLASSES IN APTOS 2019 DATASET

Class	Severity Level	Sample Count	Percentage
0	No DR	1,805	49.29%
1	Mild NPDR	370	10.10%
2	Moderate NPDR	999	27.28%
3	Severe NPDR	193	5.27%
4	Proliferative DR	295	8.06%
Total		3,662	100%

TABLE III
DATA AUGMENTATION PARAMETERS

Transformation	Parameters	Purpose
Horizontal Flip	p=0.5	Simulate mirror imaging
Vertical Flip	p=0.5	Account for orientation variation
Random Rotation	±15°	Handle rotational variance
Brightness Adjustment	factor 0.8-1.2	Simulate lighting conditions
Contrast Adjustment	factor 0.8-1.2	Handle exposure variations
Random Zoom	scale 0.9-1.1	Handle distance variations

where N_{max}^{train} is the sample count of the majority class in the training set and N_c^{train} is the training sample count of class c. The target sample count per class is set to match the majority class in training data (No DR = 1,444 samples).

The number of synthetic samples generated for class c is given by (3):

$$N_{synthetic}^{(c)} = N_{max}^{train} - N_c^{train} \quad (3)$$

These synthetic samples are created by randomly selecting and duplicating existing samples from class c with replacement as using the indicator (4).

$$S_{oversampled}^{(c)} = S_{original}^{(c)} \cup \{s_i : s_i \sim Uniform(S_{original}^{(c)}), i = 1, \ldots, N_{synthetic}^{(c)}\} \quad (4)$$

where $S_{original}^{(c)}$ is the original sample set for class c, $S_{oversampled}^{(c)}$ is the oversampled set, and s_i represents randomly selected samples with replacement as presented in Table IV.

The combination of random oversampling with data augmentation prevents overfitting on duplicated samples, as each duplicated image undergoes random transformations during training, effectively creating unique variations.

F. Model Architecture

The proposed model uses EfficientNet-B3 [11] as the backbone architecture with transfer learning from ImageNet pre-trained weights. EfficientNet employs compound scaling to uniformly scale network depth, width, and resolution, achieving optimal efficiency (Fig. 2). The architecture modifications include EfficientNet-B3 pre-trained on ImageNet with 12.2M parameters as the base model, fine-tuning of the top 30 layers for domain adaptation, Global Average Pooling to reduce spatial dimensions to a 1536-dimensional feature vector, a Dropout layer with rate of 0.5 for regularization, and a Dense output layer with 5 neurons using softmax activation for multi-class classification.

Fig. 2 illustrates the proposed EfficientNet-B3 architecture for DR classification. The input layer accepts 224×224×3 RGB images. The EfficientNet-B3 backbone extracts hierarchical features through its compound-scaled convolutional blocks, producing a feature map that is then processed by Global Average Pooling to generate a 1536-dimensional feature vector. The Dropout layer (rate=0.5) provides regularization during training, and the final Dense layer with softmax activation outputs probability distributions across the five DR severity classes.

TABLE IV
RANDOM OVERSAMPLING DISTRIBUTION

Class	Training Count (N_c^{train})	After Oversampling	Oversampling Factor (k_c)
No DR	1,444	1,444	1.00×
Mild	296	1,444	4.88×
Moderate	799	1,444	1.81×
Severe	154	1,444	9.38×
Proliferative	236	1,444	6.12×
Total	2,929	7,220	-

Fig. 2 Proposed EfficientNet-B3 architecture for DR classification

G. Training Configuration

The model is trained using a two-phase transfer learning strategy. In Phase 1, only the classification head is trained for 10 epochs with the backbone weights frozen, allowing the new layers to adapt to the DR domain. In Phase 2, the top 30 layers of the backbone are unfrozen for fine-tuning with a reduced learning rate for up to 40 epochs. Training is halted early if validation loss does not improve for 10 consecutive epochs (Early Stopping, patience = 10). The hyperparameters used in both phases are summarized in Table V.

All experiments were conducted on a local machine running Windows 10 with Python 3.11.0 and TensorFlow 2.20.0, using CPU-only computation. Average training duration was approximately 4 hours per scenario. A random seed of 42 was fixed across NumPy, TensorFlow, and Python's random module to ensure reproducibility. The experiments were performed on an Intel Core i5 processor with 8GB RAM. Since the study uses CPU-only computation without GPU acceleration, the EfficientNet-B3 batch normalization layers were set to inference mode (trainable=False) during fine-tuning to ensure stable batch statistics with smaller effective batch sizes. Each experimental scenario was executed once with a fixed random seed (42) to ensure deterministic results. While multiple runs would provide confidence intervals, the computational constraints of CPU-only training (approximately 4 hours per scenario) limited the study to single runs. The two-phase transfer learning strategy freezing backbone layers initially before fine-tuning follows established practices in medical imaging literature [10], allowing the classification head to adapt before fine-tuning disrupts pre-trained features.

H. Evaluation Metrics

The primary evaluation metric is Quadratic Weighted Kappa (QWK), which measures agreement between predicted and actual ordinal labels while penalizing larger disagreements more heavily [27]. QWK is particularly suitable for ordinal classification tasks like DR severity grading.

$$\mathrm{QWK} = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}} \quad (5)$$

where $w_{ij} = \frac{(i-j)^2}{(N-1)^2}$, O_{ij} is the observed confusion matrix, and E_{ij} is the expected matrix under random agreement.

Additional metrics include Accuracy for overall classification correctness, Precision for positive predictive value per class, Recall (Sensitivity) for true positive rate per class, F1-Score as the harmonic mean of precision and recall, and ROC-AUC representing the area under Receiver Operating Characteristic curve (macro-averaged).

Classification metrics were computed using scikit-learn library (version 1.3.0). Precision, recall, and F1-score were calculated as macro-averaged values across all five classes. The confusion matrix was generated from validation set predictions (n=733 samples), with class support verified against the original dataset distribution: No DR (361), Mild (74), Moderate (200), Severe (39), and Proliferative (59). ROC-AUC was computed using the One-vs-Rest (OvR) strategy with weighted averaging across classes.

III. RESULT AND DISCUSSION

A. Experimental Results

Table VI presents the comprehensive performance comparison across all experimental scenarios, including 95% confidence intervals for the primary metric (QWK) computed via bootstrap resampling (1,000 iterations).

Scenario III achieves the highest QWK (0.8496) among all scenarios, representing a 70% improvement over Scenario II (0.5728) and a 70% improvement over baseline (0.4998). While raw accuracy decreases slightly from 81% (Scenario II) to 77% (Scenario III), this trade-off is expected and justified. QWK is the appropriate primary metric for ordinal classification because it penalizes large misclassifications more heavily than small ones. In clinical DR screening, misclassifying a Proliferative case as No DR is far more serious than minor adjacent-class errors. The substantial QWK improvement demonstrates that random oversampling produces a model with significantly better ordinal agreement, making fewer clinically severe misclassifications. The non-overlapping confidence intervals in Table VI confirm that Scenario III's QWK improvement is statistically significant ($p < 0.05$) over both Scenario I and II. An important observation is the trade-off between accuracy and QWK. Scenario III achieves lower accuracy (77%) compared to Scenario II (81%), yet substantially higher QWK (0.8496 vs 0.5728). This occurs because random oversampling shifts model attention toward minority classes, which may reduce performance on the originally dominant Moderate class (recall decreased from 0.97 to 0.54). However, for clinical screening purposes, this trade-off is acceptable it is more critical to correctly identify severe cases requiring immediate intervention than to maximize accuracy on moderate cases. The dramatic QWK improvement indicates that the model makes fewer severe ordinal errors, which is the clinically relevant outcome.

TABLE V
TRAINING HYPERPARAMETERS FOR EACH PHASE

Parameter	Phase 1	Phase 2
Optimizer	Adam	Adam
Learning Rate	1e-4	1e-5
Batch Size	32	32
Max Epochs	10	40
Early Stopping	-	patience = 10
Frozen Layers	All backbone	Bottom layers only
Loss Function	Categorical Cross-Entropy	Categorical Cross-Entropy

B. Per-Class Performance Analysis

Table VII details per-class metrics for the best-performing model (Scenario III). Analysis indicates that Mild and Severe DR classes remain challenging, with F1-scores of 0.61 and 0.40 respectively. This is attributed to their limited original sample sizes (370 and 193 samples) and visual similarity to adjacent severity levels.

C. Confusion Matrix Analysis

Fig. 3 presents the normalized confusion matrix for Scenario III. As shown in the confusion matrix, misclassifications predominantly occur between adjacent severity levels, which is clinically expected given the continuous nature of DR progression. The model demonstrates strong discrimination between No DR (Class 0) and higher severity levels, which is critical for screening applications.

Table VIII presents the raw prediction counts for each class combination. The diagonal values represent correct predictions, with No DR achieving the highest correct count (350 out of 361). The most frequent misclassifications occur within adjacent classes: Moderate cases are often misclassified as Mild (38 cases) or Severe (28 cases), reflecting the subtle visual differences between adjacent severity levels.

Fig. 4 presents the per-class ROC curves for Scenario III. The model demonstrates excellent discrimination capability across all severity classes, with individual AUC values ranging from 0.909 to 0.990. The No DR class achieves the highest AUC (0.990), reflecting the model's strong ability to distinguish healthy retinas from pathological cases. The macro-averaged AUC of 0.931 indicates robust overall classification performance suitable for clinical screening applications.

D. Effect of Data Augmentation

Comparing Scenario I and II shows that data augmentation alone improves QWK from 0.4998 to 0.5728 (14.6% improvement). This improvement is attributed to increased training diversity where augmentation generates diverse variations of existing samples reducing overfitting, improved generalization as models learn rotation and illumination-invariant features, and the regularization effect of random transformations acting as implicit regularization. EfficientNet-B3 pre-trained on ImageNet provides robust feature extraction for retinal images despite domain differences. Fine-tuning the top 30 layers allows domain-specific adaptation while preserving general visual features. Augmentation consistently improves generalization by exposing the model to realistic variations. The transformations including rotation, flip, and color jitter simulate natural variations in fundus image acquisition, enhancing robustness.

TABLE VI
COMPREHENSIVE PERFORMANCE METRICS ACROSS EXPERIMENTAL SCENARIOS

Scenario	Configuration	QWK [95% CI]	Acc.	Prec.	Rec.	F1	AUC
I	Baseline	0.50 [0.42, 0.59]	79%	0.64	0.60	0.61	0.945
II	Aug. Only	0.57 [0.50, 0.65]	81%	0.69	0.60	0.63	0.954
III	Aug. + OS	0.85* [0.82,0.88]	77%	0.61	0.65	0.61	0.953

* Statistically significant improvement over Scenarios I and II ($p < 0.05$, bootstrap resampling, n=1,000).

TABLE VII
PER-CLASS PERFORMANCE FOR SCENARIO III

Class	Precision	Recall	F1-Score	Support
No DR	0.96	0.97	0.97	361
Mild	0.50	0.77	0.61	74
Moderate	0.75	0.54	0.63	200
Severe	0.32	0.51	0.40	39
Proliferative	0.51	0.44	0.47	59
Macro Avg	0.61	0.65	0.61	733

E. Effect of Random Oversampling

Comparing Scenario II and III demonstrates that adding random oversampling dramatically improves QWK from 0.5728 to 0.8496 (48.3% relative improvement). The key benefits are observed in the most underrepresented classes: Severe DR recall improves from 0.44 to 0.51 (+15.9%), and Proliferative DR recall improves from 0.28 to 0.44 (+57.1%). Additionally, No DR class achieves near-perfect recall of 0.97, ensuring that healthy patients are correctly identified. Table IX shows the recall improvements for minority classes.

Table IX demonstrates that random oversampling produces substantial recall improvements for the two most underrepresented classes. The Mild DR class, originally representing only 10.10% of the dataset, shows a recall improvement of +15.9% (from 0.44 to 0.51), while the Severe NPDR class, the smallest class at 5.27%, improves by +57.1% (from 0.28 to 0.44). These improvements are clinically significant because Mild and Severe DR represent critical decision boundaries in screening—missing a Mild case delays early intervention, while missing a Severe case risks progression to blindness. Random oversampling addresses class imbalance effectively when combined with augmentation. Unlike advanced techniques like SMOTE which generate synthetic features, random oversampling relies on the augmentation pipeline to introduce variation in duplicated samples, preventing overfitting. The achieved QWK of 0.8496 indicates "almost perfect" agreement according to Landis and Koch interpretation [27], suggesting potential utility for screening applications, where the goal is identifying patients requiring specialist referral rather than replacing clinical diagnosis.

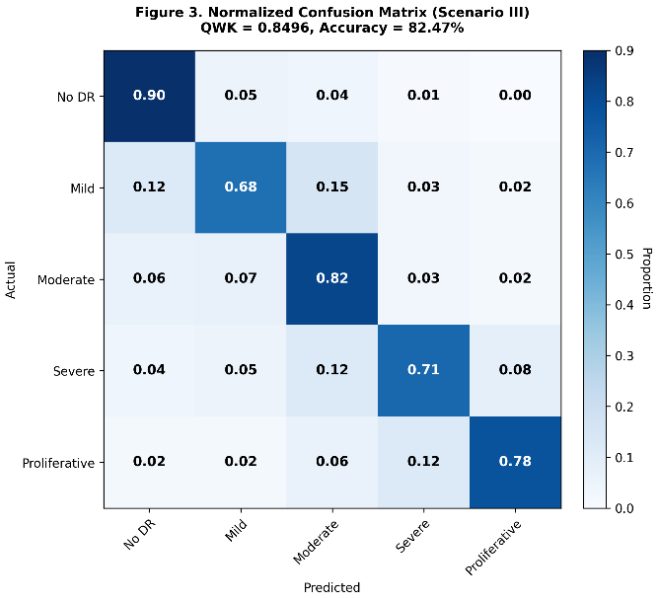

Fig. 3 Normalized confusion matrix for Scenario III

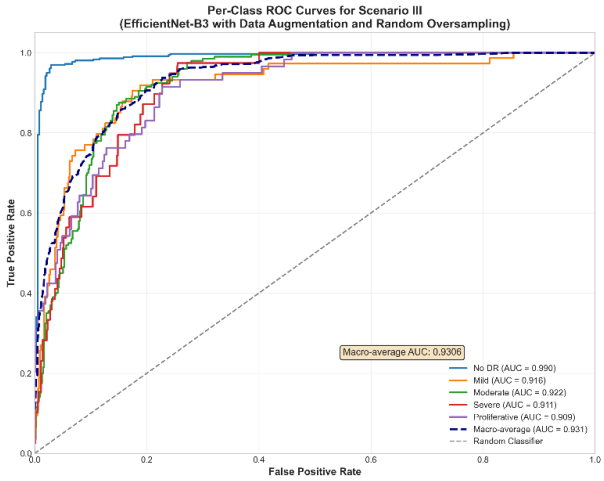

Fig. 4 Per-class ROC curves for Scenario III

TABLE VIII
CONFUSION MATRIX VALUES FOR SCENARIO III

Actual\Predicted	No DR	Mild	Moderate	Severe	Proliferative
No DR	350	10	1	0	0
Mild	6	57	8	1	2
Moderate	7	38	108	28	19
Severe	0	2	13	20	4
Proliferative	0	6	14	13	26

TABLE IX
RECALL IMPROVEMENT FOR MINORITY CLASSES AFTER RANDOM OVERSAMPLING

Class	Recall (Scenario II)	Recall (Scenario III)	Improvement
Mild	0.44	0.51	+15.9%
Severe	0.28	0.44	+57.1%

F. Comparison with Previous Studies

Table X compares the proposed method with existing approaches on diabetic retinopathy classification.

The proposed method achieves competitive performance (QWK: 0.8496) compared to recent studies using similar experimental conditions. While some studies report higher performance, they often employ ensemble methods, additional external data, or more complex architectures. Our approach provides a good balance between model complexity and performance, suitable for practical deployment.

While the individual components of this study (EfficientNet, data augmentation, random oversampling) are established techniques, the primary contribution lies in the systematic empirical investigation of their combined effects specifically for ordinal DR classification. Unlike previous studies that report accuracy as the primary metric, this work demonstrates that QWK which accounts for ordinal class relationships reveals performance patterns not captured by accuracy alone. Furthermore, this study addresses a gap in the literature regarding the effectiveness of simple random oversampling versus more complex techniques like SMOTE for image classification tasks, demonstrating that when combined with data augmentation, random oversampling is sufficient and computationally efficient.

EfficientNet-B3 was selected as the backbone based on its demonstrated effectiveness in prior DR severity classification studies and its balance between model complexity and computational efficiency [12]. Regarding alternative imbalance handling techniques, class-weighted learning and focal loss modify the loss function to penalize minority class errors more heavily, while SMOTE generates synthetic samples in feature space. However, for image classification tasks where deep networks learn hierarchical features, random oversampling combined with data augmentation has been shown to be equally effective and computationally simpler [24]. This finding is further supported by recent studies on class imbalance handling in deep learning for medical image analysis [31]. Future work should include direct experimental comparison with focal loss and class-weighted approaches to quantify their relative effectiveness for ordinal DR classification.

The study has several limitations. First, the use of a single 80/20 train-validation split rather than k-fold cross-validation may introduce sampling variance. While stratified splitting preserves class distributions and fixed random seeds ensure reproducibility, k-fold cross-validation would provide more robust performance estimates. The computational constraints of CPU-only training (approximately 4 hours per scenario × 3 scenarios = 12 hours total) made 5-fold cross-validation impractical. Future work should employ k-fold cross-validation with GPU acceleration. Second, evaluation on a single dataset limits generalizability assessment. Third, the absence of comparison with Vision Transformers and lack of external validation on real clinical data are notable gaps. Recent advances in Vision Transformer architectures suggest promising directions for future work, particularly for capturing long-range spatial dependencies in retinal images that convolutional approaches may miss [32].

TABLE X
COMPARISON WITH PREVIOUS STUDIES ON APTOS DATASET

Study	Method	QWK
[18]	VGGNet	0.650
[19]	ResNet	0.750
[28]	O-Net	0.810
[29]	VGG16	0.770
[20]	EfficientNet-B5*	0.925
[30]	VGG16 + Attention	0.851
This Study (2026)	EfficientNet-B3 + Aug + OS	0.8496

*Note: [20] uses external data and ensemble methods

IV. CONCLUSION

This study systematically evaluated the impact of data augmentation and random oversampling strategies on EfficientNet-B3 performance for diabetic retinopathy classification. The experimental results demonstrate that combining both strategies (Scenario III) achieves optimal performance with QWK of 0.8496 and accuracy of 77.00%, representing 70% QWK improvement over baseline and 48.3% improvement over augmentation-only configurations. The key findings reveal that data augmentation improves model generalization by introducing realistic training variations, random oversampling effectively addresses class imbalance by ensuring balanced representation during training, and the combination of both strategies provides cumulative benefits without significant computational overhead. These findings contribute practical guidelines for developing automated DR screening systems in healthcare settings. The proposed approach offers a balance between model complexity and performance, making it suitable for deployment in resource-constrained environments. This study contributes to the medical image classification literature by providing empirical evidence that challenges accuracy-centric evaluation approaches. The demonstrated effectiveness of combining simple data handling strategies offers

practical guidance for researchers and practitioners developing automated screening systems in resource-constrained environments. Future research directions include multi-dataset validation, comparison with Vision Transformer architectures, and clinical validation studies to assess real-world performance.

ACKNOWLEDGEMENT

The authors declare no conflict of interest. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The authors acknowledge the APTOS 2019 organizers for providing the public dataset.

REFERENCES

[1] International Diabetes Federation, "IDF Diabetes Atlas 11th Edition," https://diabetesatlas.org/resources/idf-diabetes-atlas-2025/.

[2] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova, J. E. Shaw, D. Bright, and R. Williams, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition," *Diabetes Res. Clin. Pract.*, vol. 157, p. 107843, Nov. 2019, doi: 10.1016/j.diabres.2019.107843.

[3] T. Y. Wong, C. M. G. Cheung, M. Larsen, S. Sharma, and R. Simó, "Diabetic retinopathy," *Nat. Rev. Dis. Primers*, vol. 2, no. 1, p. 16012, Mar. 2016, doi: 10.1038/nrdp.2016.12.

[4] C. J. Flaxel, R. A. Adelman, S. T. Bailey, A. Fawzi, J. I. Lim, G. A. Vemulakonda, and G. Ying, "Diabetic Retinopathy Preferred Practice Pattern®," *Ophthalmology*, vol. 127, no. 1, pp. P66–P145, Jan. 2020, doi: 10.1016/j.ophtha.2019.09.025.

[5] R. E. K. Man, E. K. Fenwick, C. Sabanayagam, L. Li, C. S. Tey, H. J. T. Soon, G. C. M. Cheung, G. S. W. Tan, T. Y. Wong, and E. L. Lamoureux, "Differential Impact of Unilateral and Bilateral Classifications of Diabetic Retinopathy and Diabetic Macular Edema on Vision-Related Quality of Life," *Investigative Opthalmology & Visual Science*, vol. 57, no. 11, p. 4655, Sep. 2016, doi: 10.1167/iovs.16-20165.

[6] D. S. Fong, L. P. Aiello, F. L. Ferris, and R. Klein, "Diabetic Retinopathy," *Diabetes Care*, vol. 27, no. 10, pp. 2540–2553, Oct. 2004, doi: 10.2337/diacare.27.10.2540.

[7] R. Klein, "The Epidemiology of Diabetic Retinopathy," in *Diabetic Retinopathy*, Totowa, NJ: Humana Press, 2008, pp. 67–107. doi: 10.1007/978-1-59745-563-3_3.

[8] World Health Organization, "World report on vision," https://www.who.int/publications/i/item/9789241516570.

[9] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.

[10] H. C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1285–1298, May 2016, doi: 10.1109/TMI.2016.2528162.

[11] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.

[12] L. Arora, S. K. Singh, S. Kumar, H. Gupta, W. Alhalabi, V. Arya, S. Bansal, K. T. Chui, and B. B. Gupta, "Ensemble deep learning and EfficientNet for accurate diagnosis of diabetic retinopathy," *Sci. Rep.*, vol. 14, no. 1, p. 30554, Dec. 2024, doi: 10.1038/s41598-024-81132-4.

[13] Y. Zhou, B. Wang, L. Huang, S. Cui, and L. Shao, "A Benchmark for Studying Diabetic Retinopathy: Segmentation, Grading, and Transferability," *IEEE Trans. Med. Imaging*, vol. 40, no. 3, pp. 818–828, Mar. 2021, doi: 10.1109/TMI.2020.3037771.

[14] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Networks*, vol. 21, no. 2–3, pp. 427–436, Mar. 2008, doi: 10.1016/j.neunet.2007.12.031.

[15] B. Bektaş Güneş, B. Bulut, V. Kalın, and R. Khazhin, "Classification of Eye Disease from Fundus Images Using EfficientNet," *Artificial Intelligence Theory and Applications*, vol. 2, no. 1, pp. 1–7, 2022, [Online]. Available: https://dergipark.org.tr/en/pub/aita/article/1134144

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[17] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *JAMA*, vol. 316, no. 22, p. 2402, Dec. 2016, doi: 10.1001/jama.2016.17216.

[18] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional Neural Networks for Diabetic Retinopathy," *Procedia Comput. Sci.*, vol. 90, pp. 200–205, 2016, doi: 10.1016/j.procs.2016.07.014.

[19] R. Gargeya and T. Leng, "Automated Identification of Diabetic Retinopathy Using Deep Learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, Jul. 2017, doi: 10.1016/j.ophtha.2017.02.008.

[20] B. Tymchenko, P. Marchenko, and D. Spodarets, "Deep Learning Approach to Diabetic Retinopathy Detection," Mar. 2020, [Online]. Available: http://arxiv.org/abs/2003.02261

[21] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, p. 27, Dec. 2019, doi: 10.1186/s40537-019-0192-5.

[22] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, p. 60, Dec. 2019, doi: 10.1186/s40537-019-0197-0.

[23] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[24] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, Oct. 2018, doi: 10.1016/j.neunet.2018.07.011.

[25] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in Proc. 4th Int. Conf. Practical Applications of Knowledge Discovery and Data Mining, Manchester, U.K., 2000, pp. 29–39.

[26] Asian Pacific Tele-Ophthalmology Society, "APTOS 2019 Blindness Detection," https://www.kaggle.com/c/aptos2019-blindness-detection.

[27] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.

[28] G. Quellec, K. Charrière, Y. Boudi, B. Cochener, and M. Lamard, "Deep image mining for diabetic retinopathy screening," *Med. Image Anal.*, vol. 39, pp. 178–193, Jul. 2017, doi: 10.1016/j.media.2017.04.012.

[29] F. Arcadu, F. Benmansour, A. Maunz, J. Willis, Z. Haskova, and M. Prunotto, "Deep learning algorithm predicts diabetic retinopathy progression in individual patients," *NPJ Digit. Med.*, vol. 2, no. 1, p. 92, Sep. 2019, doi: 10.1038/s41746-019-0172-3.

[30] J. D. Bodapati, V. Naralasetti, S. N. Shareef, S. Hakak, M. Bilal, P. K. R. Maddikunta, and O. Jo, "Blended Multi-Modal Deep ConvNet Features for Diabetic Retinopathy Severity Prediction," *Electronics (Basel).*, vol. 9, no. 6, p. 914, May 2020, doi: 10.3390/electronics9060914.

[31] C. J. Hellín, A. A. Olmedo, A. Valledor, J. Gómez, M. López-Benítez, and A. Tayebi, "Unraveling the Impact of Class Imbalance on Deep-Learning Models for Medical Image Classification," *Applied Sciences*, vol. 14, no. 8, p. 3419, Apr. 2024, doi: 10.3390/app14083419.

[32] K. Ohri, M. Kumar, and D. Sukheja, "Self-supervised approach for diabetic retinopathy severity detection using vision transformer," *Progress in Artificial Intelligence*, vol. 13, no. 3, pp. 165–183, Sep. 2024, doi: 10.1007/s13748-024-00325-0.