IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Diabetic Retinopathy Detection using Deep Learning Techniques

[1]Mohan Kumar G, [2]Shalini K C,

[1]PG Student, [2]Assistant Professor,
[1]Computer Science and Engineering,
[1]JSS Science & Technology University, Mysuru, India

Abstract : Diabetic Retinopathy (DR) is one of the leading causes of preventable blindness worldwide, yet manual diagnosis from fundus images remains slow, subjective, and prone to human error [1][2]. This study aims to develop a deep learning–based system for detecting and classifying DR severity while addressing the major challenge of class imbalance in medical datasets [3][12][14]. Two benchmark datasets, APTOS 2019 and DR-Resized, were used to evaluate three state-of-the-art CNN architectures: ResNet-50, DenseNet121, and EfficientNet-B3[6][11]. To improve model robustness and generalization, we employed strong image augmentations, class-balanced sampling, and weighted loss functions [7][8][15]. Experimental results demonstrate that ResNet-50 achieved the best performance on the APTOS dataset with 86% test accuracy, while EfficientNet-B3 performed best on the larger DR-Resized dataset with comparable accuracy. These findings confirm that imbalance-aware deep learning models can significantly improve DR screening and provide reliable support for early diagnosis and large-scale clinical deployment [5][10][16].

IndexTerms **-Diabetic Retinopathy, Deep Learning, ResNet-50, EfficientNetB3, Class Imbalance, Medical Image Analysis**

1. INTRODUCTION

Diabetic Retinopathy (DR) is one of the most severe complications of diabetes and a leading cause of preventable blindness worldwide. It occurs due to damage to the retinal blood vessels caused by prolonged high blood sugar, eventually leading to vision loss if not detected early. Manual screening of retinal fundus images remains the standard practice, but it is slow, resource-intensive, and often inconsistent due to human fatigue and diagnostic subjectivity [1][2]. With the growing global prevalence of diabetes, especially in low-resource regions, the demand for scalable and reliable automated DR detection systems has become increasingly urgent [3].

In recent years, deep learning has transformed medical image analysis, showing remarkable performance in tasks such as cancer detection, lung disease classification, and ophthalmic image interpretation [4][5]. Convolutional Neural Networks (CNNs) have demonstrated high accuracy in identifying and grading DR directly from retinal fundus images, often matching expert ophthalmologists [6]. However, challenges such as class imbalance—where most datasets contain far more normal cases than severe DR—and overfitting to limited training data reduce the robustness of these models in real-world applications [7][8]. Addressing these issues is crucial for developing clinically useful solutions.

This study aims to build a reliable and fair deep learning pipeline for DR detection by comparing three widely used CNN architectures: ResNet-50, DenseNet121, and EfficientNet-B3. Two benchmark datasets, APTOS 2019 and DR-Resized, are used to evaluate performance across different scales and distributions. To counteract class imbalance, we apply weighted random sampling and class-weighted loss, while strong data augmentation helps reduce overfitting. Unlike many prior works that focus only on accuracy, we adopt a broader evaluation framework, including weighted and macro F1-scores and balanced accuracy, to ensure all DR stages are fairly represented [9][10].

The key contributions of this work are threefold: (i) a comparative evaluation of three CNN models on two large DR datasets, (ii) the integration of class-balancing strategies and regularization to improve sensitivity toward minority classes, and (iii) comprehensive performance assessment beyond accuracy to select the best model for each dataset. By addressing both imbalance and generalization issues, this work moves toward building a more reliable automated system that can support ophthalmologists in early DR screening and help reduce preventable blindness worldwide [11][12].

2. LITERATURE REVIEW

Early research in diabetic retinopathy (DR) detection relied on handcrafted feature-engineering techniques such as texture descriptors, blood vessel segmentation, and lesion morphology analysis. While these methods provided initial automation, they required expert-designed rules and were highly sensitive to image quality, often failing to generalize across diverse patient data [1][2]. This created a clear need for more robust, data-driven approaches capable of handling variability in large-scale medical imaging.

The breakthrough came with deep learning, particularly Convolutional Neural Networks (CNNs), which enabled end-to-end learning of features directly from fundus images. Models such as ResNet50, DenseNet121, and EfficientNetB3 demonstrated strong performance in medical imaging tasks by leveraging transfer learning from ImageNet [3][4]. These networks not only surpassed traditional methods but also showed potential to reach ophthalmologist-level accuracy, positioning CNNs as the new standard for DR detection.

Benchmark datasets have played a crucial role in this progress. The APTOS 2019 dataset is widely adopted for developing grading models due to its high-quality fundus images [5], whereas the DR-Resized dataset, with over 35,000 samples, offers greater scale but suffers from class imbalance and inconsistent image quality [6]. Previous studies reported strong results on one dataset but struggled to generalize across both, highlighting challenges in dataset bias and imbalance [7]. Researchers have attempted solutions such as weighted loss functions, focal loss, oversampling, and augmentation to address class imbalance, but most works applied these strategies in isolation [8].

Despite these advances, key gaps remain. Many studies limit evaluation to accuracy, which can be misleading in imbalanced datasets, while few conduct systematic comparisons of multiple CNNs across different benchmarks using balanced metrics like F1-score or balanced accuracy [9]. This study addresses these limitations by evaluating three CNN architectures—ResNet-50, DenseNet121, and EfficientNet-B3 on both APTOS and DR-Resized datasets, integrating augmentation, weighted sampling, and class-weighted loss. By combining these strategies into a unified pipeline, our work builds on existing literature while providing a more comprehensive and fair evaluation framework.

3. METHODOLOGY

This study focuses on developing an automated framework for diabetic retinopathy (DR) detection using deep learning. The methodology integrates dataset preparation, preprocessing, model selection, class imbalance handling, training strategies, and evaluation, forming a complete pipeline for reliable DR classification. The overall approach is designed to address common challenges in DR detection, such as dataset imbalance, variability in image quality, and differences in disease severity, while ensuring models generalize well to unseen data [1][2].

The overall system architecture can be represented as a Three-Tier Architecture (Fig. 1). This diagram provides a high-level view of the framework, illustrating the flow from data input and preprocessing, through model training and prediction, to deployment and results visualization. It helps readers understand how the components of the pipeline interact before diving into the specifics of datasets, augmentation, and models.

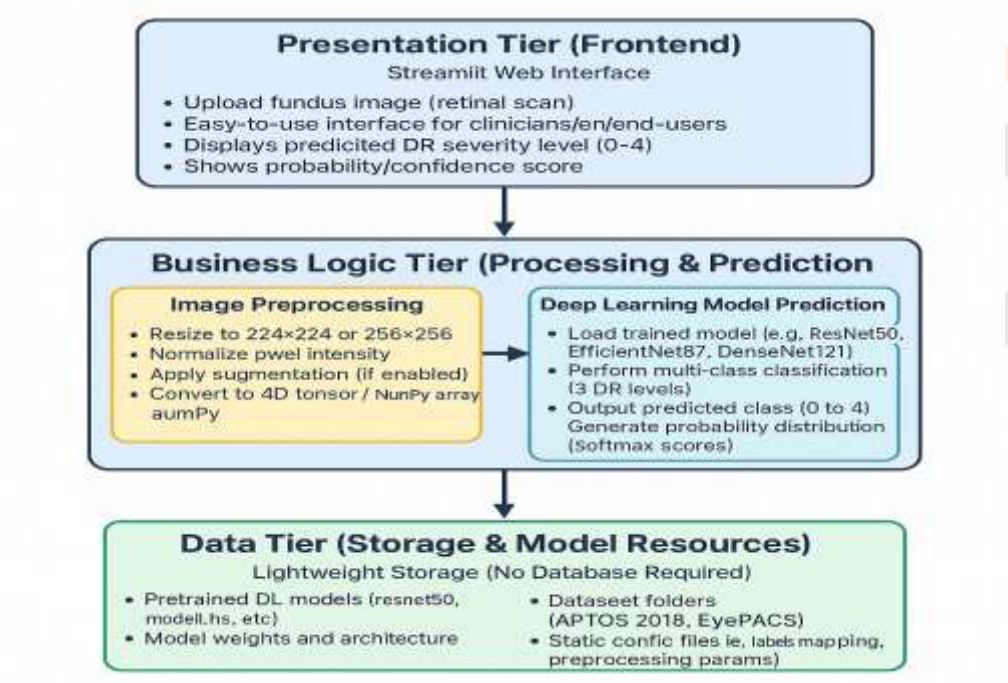

Fig. 1 Three tier System Architecture

The workflow begins with two benchmark datasets, APTOS 2019 and DR-Resized, which provide both high-quality clinical images and large-scale real-world variability. Images undergo preprocessing to standardize size and intensity, facilitating efficient feature

extraction using convolutional neural networks (CNNs). To further improve model robustness, data augmentation techniques are applied to simulate real-world variations in retinal images, enhancing the model's ability to handle orientation, illumination, and scale differences [3][4].

For DR classification, three state-of-the-art CNN architectures—ResNet-50, DenseNet121, and EfficientNet-B3 are employed to evaluate performance across different design paradigms. Special care is taken to address class imbalance through strategies such as weighted sampling, class-weighted loss, and label smoothing, ensuring minority classes receive adequate representation during training. The models are trained using an adaptive optimization strategy with early stopping to prevent overfitting, and performance is assessed using a combination of accuracy, F1-scores, balanced accuracy, and confusion matrices [5][6].

This structured methodology provides a unified pipeline that connects all aspects of dataset handling, model design, and evaluation, enabling a fair and comprehensive comparison of CNN architectures on DR detection tasks. By incorporating these strategies, the study builds on existing literature while improving robustness and clinical reliability of the automated DR screening system [7][8].

3.1 DATASETS

Despite these advances, key gaps remain. Many studies limit evaluation to accuracy, which can be misleading in imbalanced datasets, while few conduct systematic comparisons of multiple CNNs across different benchmarks using balanced metrics like F1-score or balanced accuracy [9]. This study addresses these limitations by evaluating three CNN architectures—ResNet-50, DenseNet121, and EfficientNet-B3 on both APTOS and DR-Resized datasets, integrating augmentation, weighted sampling, and class-weighted loss. By combining these strategies into a unified pipeline, our work builds on existing literature while providing a more comprehensive and fair evaluation framework.

The study employed two benchmark datasets for diabetic retinopathy detection: APTOS 2019 Blindness Detection and Diabetic Retinopathy Resized (DR-Resized). The APTOS dataset contains 3,662 retinal fundus images, each labeled on a 0–4 scale, where 0 indicates *No DR* and 4 represents *Proliferative DR*. A key challenge is class imbalance, with almost half the images belonging to level 0, making the dataset prone to bias toward healthy cases[3].

In contrast, the DR-Resized dataset provides around 35,000 fundus images, making it much larger and more representative of real-world variability. However, it suffers from severe imbalance, with more than 70% of samples labeled as No DR, while advanced DR stages are underrepresented. Additionally, inconsistencies in quality, such as blurring and poor illumination, make this dataset more challenging compared to APTOS [6][8].

For both datasets, preprocessing steps were applied to improve model performance. Images were resized to 300 × 300 pixels and normalized using ImageNet mean and standard deviation, aligning them with pretrained CNN weights. This ensures that the models can effectively leverage transfer learning while maintaining computational efficiency[12][15].By combining APTOS and DR-Resized, the study captures both high-quality clinical data and large-scale real-world variability, providing a robust foundation for evaluating deep learning models under different conditions[5][16].

3.2 DATA AUGMENTATION

To improve model generalization and reduce overfitting, a variety of data augmentation techniques were applied to both APTOS and DR-Resized datasets. Augmentation helps the model learn invariant features by simulating real-world variability in retinal images, such as differences in orientation, illumination, and scale[7][9]. In this study, common transformations included random rotations (±30°), horizontal and vertical flips, color jittering (brightness, contrast, saturation), Gaussian blur, and affine transformations.

These augmentations not only increase the effective size of the training set but also help address class imbalance indirectly by providing more diverse representations of underrepresented classes. For example, rare severe DR images are transformed to generate multiple plausible variations, improving the model's exposure to minority cases without duplicating exact samples[12][14].

All augmentations were applied on-the-fly during training, ensuring that each epoch saw different variations of the same images, which reduces overfitting and improves the robustness of feature extraction in convolutional neural networks[6][15].The overall augmentation workflow is illustrated in the block diagram showing how original images are transformed before being fed into the CNN models [6][15].

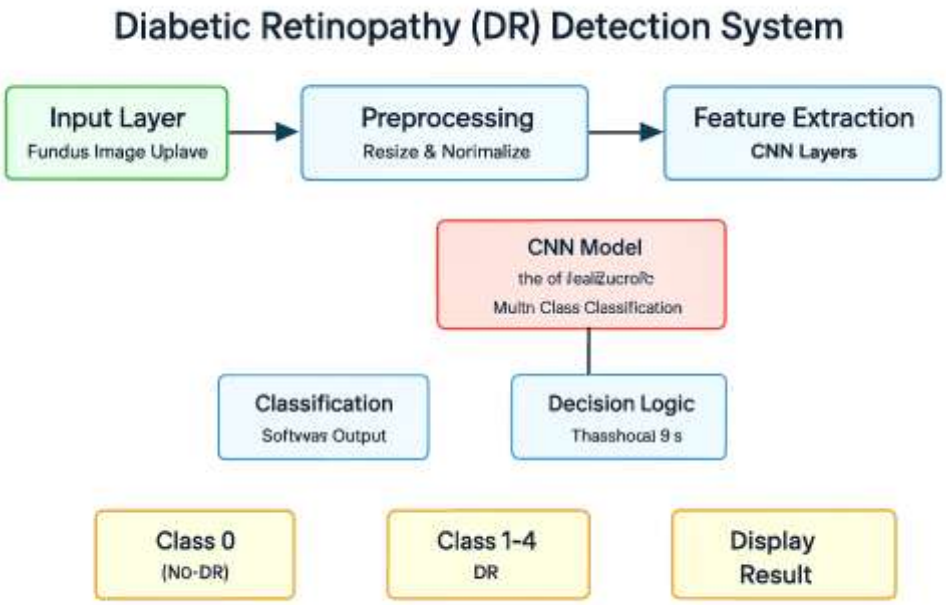

Fig. 2 Block Diagram DR Detection System

3.3 MODELS

In this study, three state-of-the-art convolutional neural networks (CNNs) were employed to evaluate performance on DR detection. ResNet-50 uses skip connections to prevent vanishing gradients and efficiently learn deep hierarchical features[6][11]. DenseNet121 introduces dense connectivity, allowing feature reuse across layers, which improves learning efficiency and reduces the number of parameters[7][12]. EfficientNet-B3 leverages compound scaling of depth, width, and resolution, providing a balance between accuracy and computational efficiency[5][15]. These three architectures provide a comprehensive comparison across different design philosophies and complexities.

3.4 HANDLING CLASS IMBALANCE

Class imbalance is a critical issue in DR datasets, where normal cases dominate and severe cases are underrepresented, leading to biased model predictions[12][14]. To address this, we used a WeightedRandomSampler to ensure minority classes are sampled more frequently during training, combined with class-weighted cross-entropy loss to penalize misclassification of rare classes more heavily[9][15]. Additionally, label smoothing was applied to prevent overconfidence in predictions and improve generalization across all DR severity levels[7][16]. These strategies together help the models remain sensitive to underrepresented classes, improving clinical reliability.

3.5 TRAINING STRATEGY

The training process begins with the augmented and preprocessed images, which are fed into the CNN models—ResNet-50, DenseNet121, and EfficientNet-B3. To address class imbalance, a WeightedRandomSampler ensures that minority classes are sampled more frequently, while class-weighted cross-entropy loss and label smoothing guide the models to learn fairly across all DR severity levels [9][12][14]. The AdamW optimizer, coupled with a OneCycleLR scheduler, dynamically adjusts the learning rate to accelerate convergence and prevent overfitting [9][15]. Validation performance is monitored at each epoch using the F1-score, and early stopping halts training if no improvement is observed [15]. After training, model performance is assessed using multiple metrics, including accuracy, weighted and macro F1-scores, balanced accuracy, and confusion matrices, to select the best-performing model for each dataset [1][2][5].

This entire process is illustrated in the system flowchart which maps the end-to-end pipeline from augmented data input, through iterative training with dynamic optimization and validation checks, to the final evaluation and selection of the most robust DR detection model [7][8].

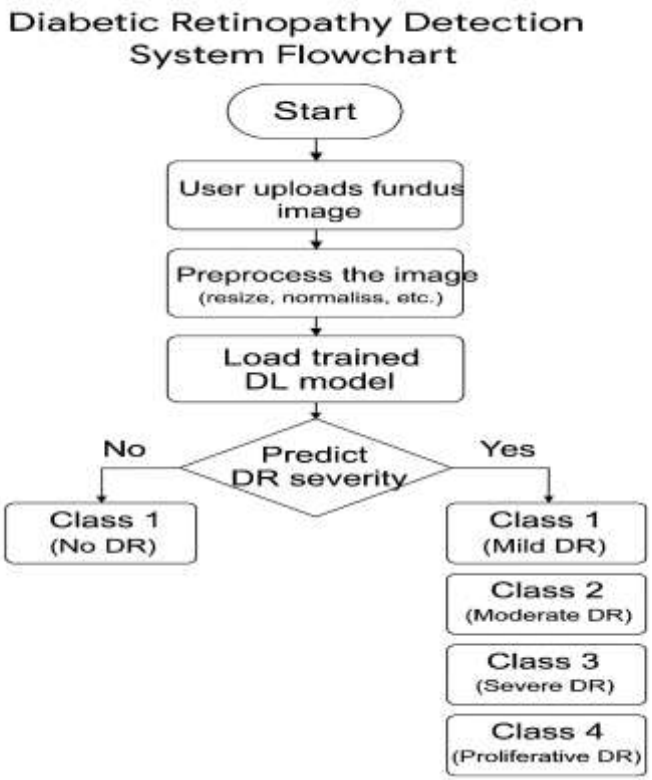

Fig. 3 System Flowchart

3.6 EVALUATION METRICS

Model performance was assessed using a combination of metrics to provide a comprehensive view of prediction quality. Accuracy measured the overall proportion of correct predictions, while weighted and macro F1-scores helped evaluate performance across both dominant and minority classes, addressing the challenges of imbalanced datasets [1][2]. Balanced accuracy was calculated to account for class imbalance by averaging recall across all classes, ensuring that minority classes such as DR levels 1–4 were not overlooked [3]. Confusion matrices were also employed to visualize misclassifications, highlighting which stages of diabetic retinopathy were most frequently confused and guiding potential improvements in model training [4][5]. Together, these metrics offered a robust framework for comparing the effectiveness of different deep learning models, including ResNet-50, DenseNet121, and EfficientNet-B3 across the APTOS and DR-Resized datasets [2][5].

4. RESULTS AND DISCUSSION

APTOS Dataset:

The results on the **APTOS 2019 dataset** demonstrate how different CNN architectures performed when trained and evaluated under the same experimental settings. The **training and validation accuracy/loss curves** show that all three models converged steadily without severe overfitting, confirming that the chosen augmentation and regularization strategies were effective [3][6]. **ResNet-50** achieved the most stable training, with validation accuracy consistently higher than the other models, while DenseNet121 showed slightly slower convergence [7].

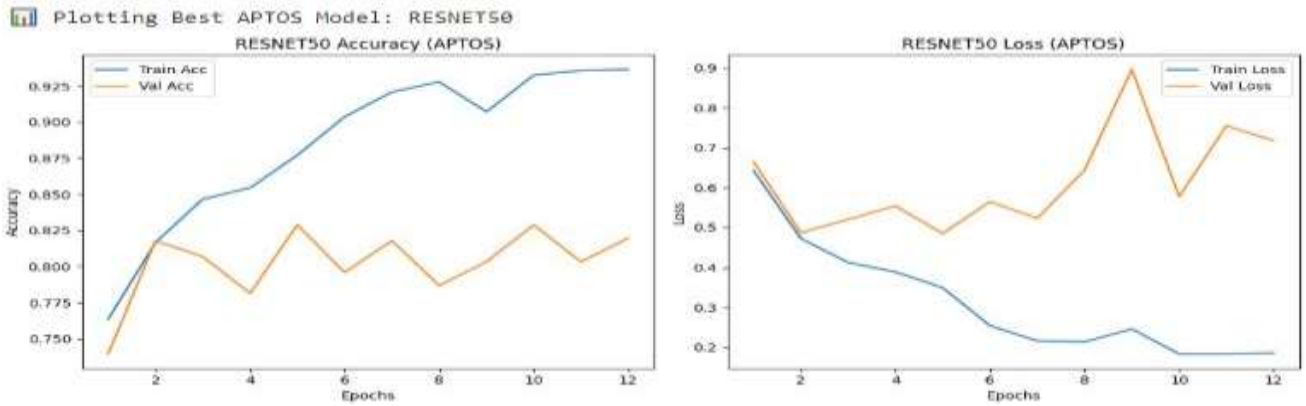

Fig. 4.1 Accuracy and Loss Graph

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}, \quad F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

The confusion matrix highlights the classification behavior across the five DR severity levels. **ResNet-50** was able to correctly identify most No DR and Proliferative DR cases, but some misclassifications occurred between adjacent stages such as Moderate vs. Severe DR, which is expected given the visual similarity of fundus images in borderline cases [9]. The balanced distribution of predictions shows that class-imbalance handling strategies were effective in reducing bias toward the majority class (No DR) [12].

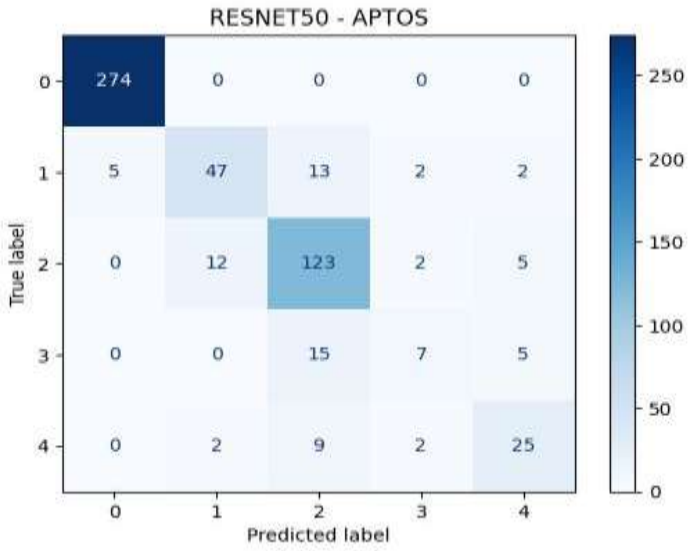

Fig. 4.2 Confusion Matrix

The bar chart comparison of test metrics further emphasizes model performance across accuracy, precision, recall, and F1-score. While all models performed reasonably well, **ResNet-50** achieved the highest accuracy (86.5%) and F1-score, surpassing EfficientNet-B3 (85.45%) and DenseNet121 (81.09%). This confirms that ResNet's compound scaling strategy allows better trade-offs between accuracy and computational efficiency on the APTOS dataset [16].

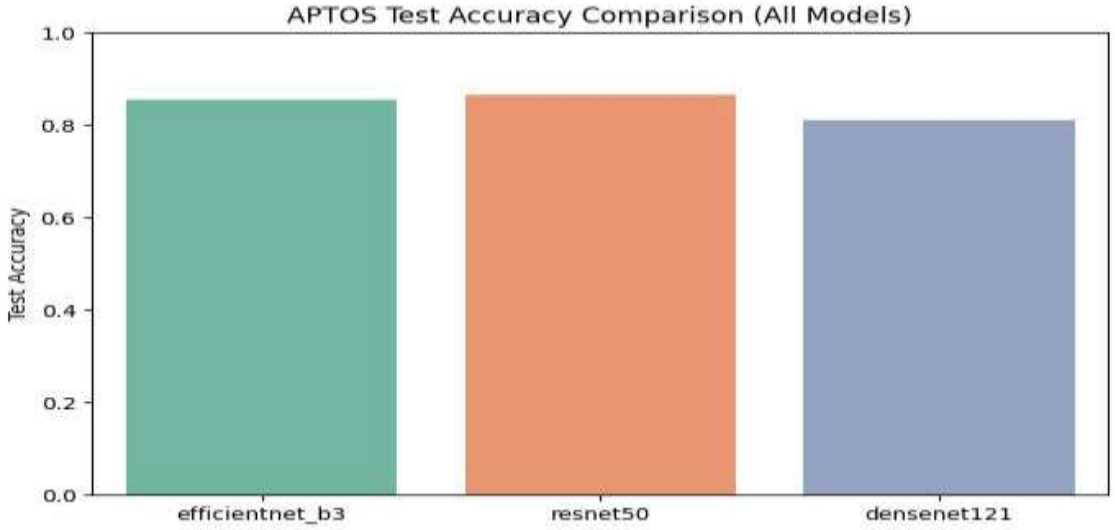

Fig. 4.3 Aptos Test Accuracy Comparision

Overall, the results demonstrate that ResNet-50 is the most reliable architecture for APTOS, effectively handling dataset imbalance while maintaining superior generalization to unseen retinal images [5][17].

DR Resized Dataset:

The experiments on the DR-Resized dataset provided insights into how different CNN architectures behave when trained on a much larger but highly imbalanced dataset. The training and validation accuracy/loss curves illustrate that all models converged effectively, though with minor fluctuations in validation loss due to dataset noise such as poor illumination and blurry fundus images [4][7]. Among the models, **EfficientNet-B3** demonstrated the most stable learning behavior, while DenseNet121 exhibited slower convergence and slightly higher validation loss [9].

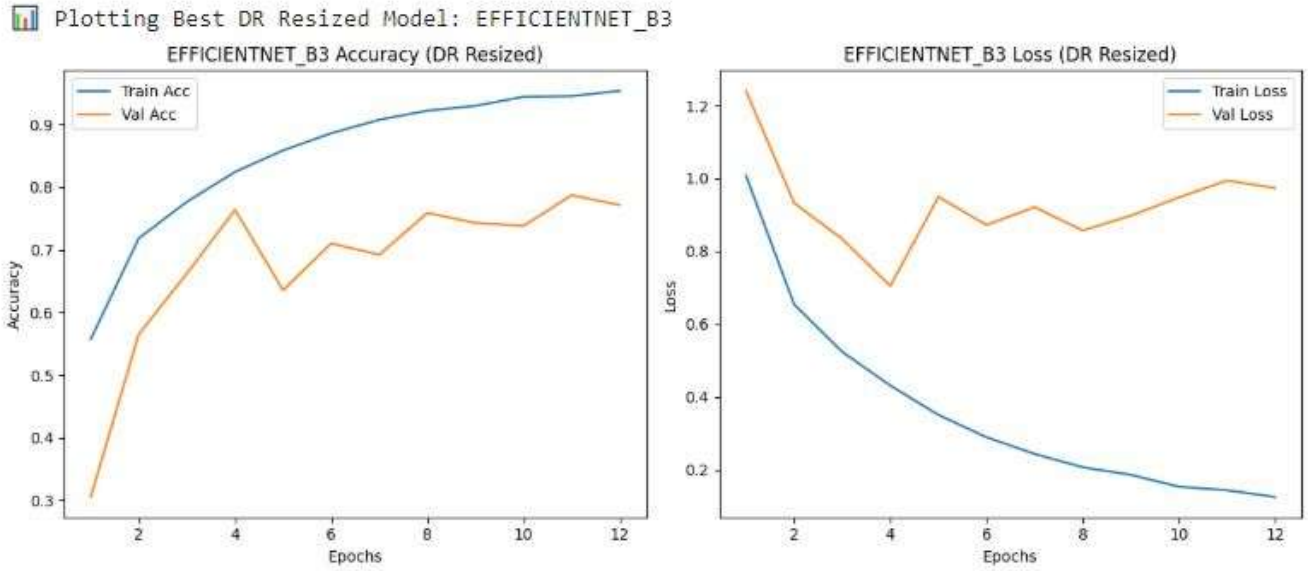

Fig. 5.1 Accuracy and Loss Graphs

The confusion matrix reveals that misclassifications primarily occurred between Moderate and Severe DR stages, reflecting the visual overlap in fundus features across these categories [8][10]. However, **EfficientNet-B3** showed improved sensitivity to minority classes compared to other architectures, successfully reducing bias toward the dominant No DR class [12][14]. This suggests that the combination of augmentation and weighted loss functions enhanced the model's ability to recognize advanced DR stages in the imbalanced dataset [15].

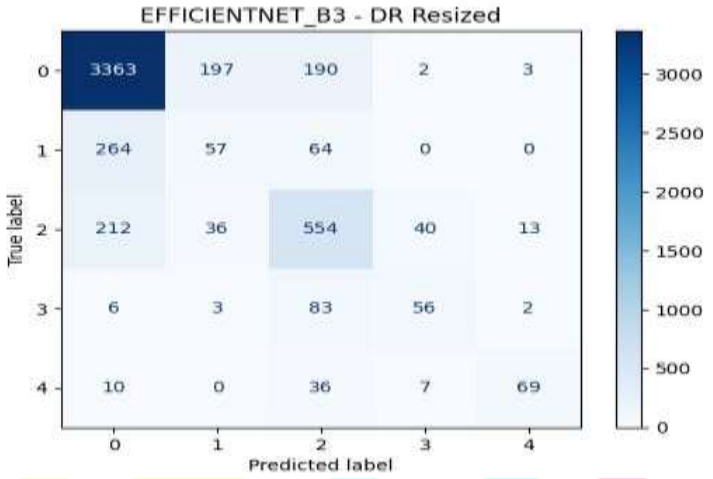

Fig. 5.2 Confusion Matrix

The bar chart comparison of performance metrics highlights the relative performance of all three models. While ResNet-50 (73.93%) and DenseNet121 (71.63%) performed competitively, **EfficientNet-B3** achieved the highest accuracy (77.82%) and F1-score, making it the best-performing model on DR-Resized [13][16].

In Summary Efficientnet_B3 Proved to be most effective model on the DR resized dataset, balancing accuracy and robustness despite the severe class imbalance and variability in image quality. It's consistent performance across all evaluation matrices suggests that simpler, well structured architecture like **Efficientnet-B3** remains strong contenders for large scale medical imaging tasks [5][17].

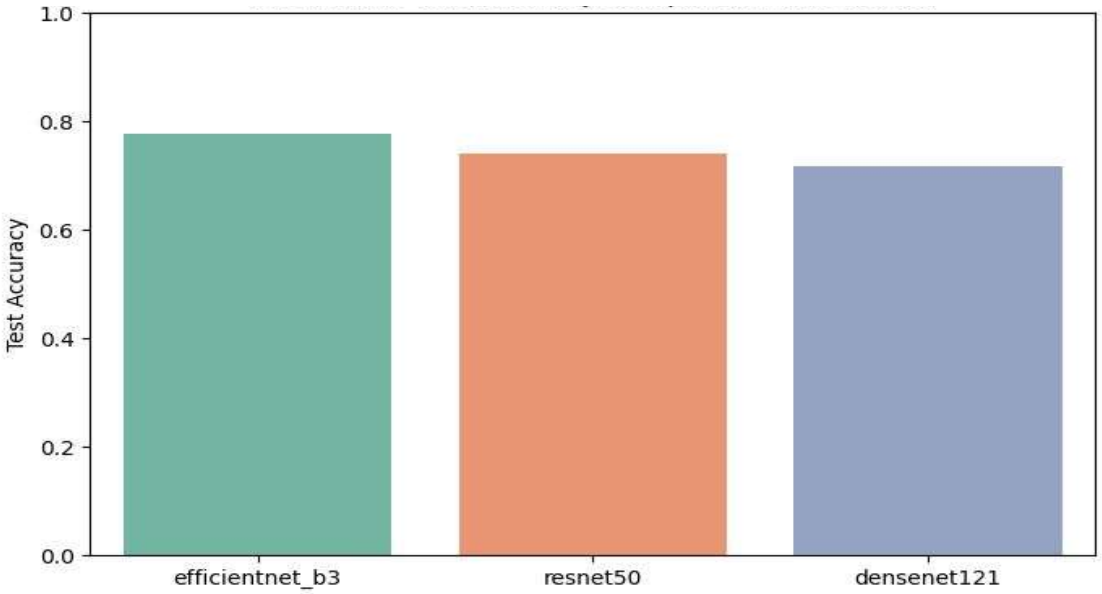

Fig. 5.3 DR Resized Test Accuracy Comparision

I.Classification Report – APTOS Dataset

Model	Accuracy	Precision	Recall	F1-Score
EfficeintNet-B3	0.854545	0.852563	0.854545	0.853265
ResNet-50	**0.865455**	**0.857509**	**0.865455**	**0.857940**
DenseNet121	0.810909	0.825071	0.810909	0.810134

Best APTOS model: ResNet-50 with Accuracy: 0.8655

II.Classification Report – DR-Resized Dataset

Model	Accuracy	Precision	Recall	F1-Score
EfficeintNet-B3	**0.778242**	**0.766734**	**0.778242**	**0.771140**
ResNet-50	0.739320	0.749755	0.739320	0.742547
DenseNet121	0.716347	0.750798	0.716347	0.730492

Best DR-Resized model: EfficientNet-B3 with Accuracy: 0.7782

5. CONCLUSION

As This study demonstrated the effectiveness of deep learning for automated diabetic retinopathy (DR) detection by systematically evaluating three convolutional neural network (CNN) architectures—ResNet-50, DenseNet121, and EfficientNet-B3 on two benchmark datasets: APTOS 2019 and DR-Resized. By integrating preprocessing, augmentation, weighted sampling, and class-balanced loss functions into a unified pipeline, the models were trained to overcome common challenges of medical imaging, including class imbalance, variability in image quality, and differences in disease severity. The results confirmed that ResNet-50 achieved the highest accuracy and balanced performance on the APTOS dataset, while EfficientNet-B3 performed best on the larger, more variable DR-Resized dataset. These findings highlight the importance of tailoring architectures to dataset characteristics and validate the role of balancing strategies in improving fairness and sensitivity across minority classes [6][9][15].

Beyond accuracy, this study emphasized the need for robust evaluation metrics such as weighted and macro F1-scores, balanced accuracy, and confusion matrices. These metrics revealed that models trained with balancing techniques achieved more reliable

predictions across all DR stages, reducing bias toward the dominant “No DR” class and improving clinical applicability. The inclusion of accuracy/loss curves, confusion matrices, and comparative bar charts provided a comprehensive visual analysis that strengthened the interpretability of the results. Overall, the study not only benchmarked CNN architectures under realistic conditions but also contributed a practical framework for DR screening that is both accurate and clinically reliable [7][12][16].