iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VI **Month of publication:** June 2025

DOI: https://doi.org/10.22214/ijraset.2025.72624

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VI June 2025- Available at www.ijraset.com

Comparative Evaluations of Deep Learning Models for Diabetic Retinopathy Classification

Shreya Basu[1], Abhijit Banerjee[2]
[1]*Master of Computer Application, Department of Computer Science and Application, University of North Bengal, Raja Rammohunpur, Bagdogra, Bairatisal, West Bengal-734014*
[2]*State aided College Teacher, Department of Computer Science, New Alipore College, University of Calcutta, Block-L, New Alipore, Kolkata-700053*

Abstract: ***This study aims to find the best fit algorithm for well-known disease Retinopathy caused by diabetes through a Deep Learning approach on two datasets. Methods: Four Deep Learning methods are employed on the datasets of Diabetic Retinopathy. The methods applied on the grayscale image files. Here, the image files from the dataset are divided into three parts: training- 80%, validation test -10% and test data-10%. We use the following preprocessing techniques to improve model generalization and address issues caused by sparse and unbalanced datasets: resize (aspect ratio), crop, normalize, augment (rotate,zoom,flipping). These techniques replicate realistic variations in retinal images and lessen overfitting. Our goal is to have clean, limited noise and test data to evaluate the model accurately. The dataset has been trained using the following methods: VGG16, ResNet50, InceptionV3 and EfficientNet correspondingly. Findings: From the above methods on the dataset-1, accuracy found from ResNet50 stood highest and on the dataset-2, accuracy found from Inception-v3 stood highest among the others. Novelty: To test generalizability and robustness, this study compares four significant convolutional neural network architectures for diabetic retinopathy classification across two heterogeneous retinal imaging datasets. Our work explores cross-dataset performance, model efficiency and clinically important metrics, providing useful insights for real-world deployment, in contrast to previous studies that were restricted to single-dataset evaluations.***
Index Terms: Diabetic retinopathy, fundus images, deep learning, convolution neural network, accuracy

I. INTRODUCTION

Diabetic Retinopathy (DR), considered as the vital pathology of vision impairment and blindness throughout the globe. These particularly affect working groups of people basically aged and in few cases juvenile. It is a microvascular impediment of prolonged diabetes mellitus. The characterization is carried out by progressive damage to the blood vessels of the human retina that ultimately leads to retinal ischemia and neovascularization. Time ahead detection and thereby intervention are critical to prevent the irreversible visual impairment. Traditional methods of DR screening such as “Fundus” photography and manual grading by ophthalmologists are rigorous, time-consuming and to a variety of observations. The evolution of artificial intelligence (AI) with machine learning (ML)and deep learning (DL) has brought breakthroughs in the field of medical imaging. Convolutional Neural Networks (CNNs) have shown great promise in automating DR diagnosis by learning hierarchical representations directly from retinal images. The groundwork for AI diagnosis began with classical ML models such as Support Vector Machines and Random Forests in the early 2000s that relied heavily on handcrafted features like texture, color and vascular patterns (Niemeijer et al., 2007).

The first breakthrough using deep learning occurred in 2015 when Gulshan et al. demonstrated a DL model that achieved sensitivity and specificity on part with ophthalmologists using a large dataset of retinal images. This work was built on CNN architectures such as AlexNet (2012), which had already transformed image classification tasks in computer vision. VGGNet (Simonyan & Zisserman, 2014) further deepened network structures, enabling improved performance on high-resolution medical images. Subsequently, ResNet (He et al., 2016) introduced residual connections, allowing much deeper networks to be trained effectively and has become a foundation model for medical image analysis. Inception networks (Szegedy et al., 2015, 2016) introduced multi-scale convolutions to capture diverse features from retinal structures. InceptionV3 showed significant performance gains in classification and localization tasks. More recently, EfficientNet (Tan & Le, 2019) proposed a novel compound scaling approach that balances network depth, width, and resolution, making it attractive for mobile and resource-constrained applications in tele-ophthalmology.

Recent literature (2020–2025) has further explored hybrid models, ensemble learning, attention mechanisms, and self-supervised learning to enhance diagnostic accuracy and generalizability. Studies such as Li et al. (2021) and Zhao et al. (2023) applied attention-based CNNs to focus on wounded areas, thereby improving explainability and trust in DL-based systems. Moreover, federated learning and privacy-preserving training paradigms have emerged to address concerns of data privacy in medical applications (Yang et al., 2020). Despite these advancements, challenges remain. Differences in image acquisition protocols, variability in disease severity across populations, and dataset imbalance continue to hinder generalization. Therefore, comparative analysis of different DL architectures under uniform experimental conditions is essential to guide practical deployment in real-world screening settings. This work presents a crude and critical comparative study of four popular CNN architectures—VGG16, ResNet50, InceptionV3 and EfficientNet—on two publicly available DR datasets. The focus is to identify the most suitable model based on classification accuracy, convergence behavior and generalization ability, thereby contributing to evidence-based model selection in automated DR detection.

II. METHODOLOGY AND EXPERIMENTAL DESIGN

The methodology adopted in this study follows a systematic pipeline for developing and evaluating deep learning models for Diabetic Retinopathy (DR) classification. The workflow is illustrated following:

A. Process

1) Step 1: Dataset Acquisition

Fundus image datasets (Dataset 1: 1,750 grayscale images; Dataset 2: 15,000 color images) are collected from publicly available open sources.

2) Step 2: Preprocessing

- Convert to grayscale (Dataset 1 only)
- Resize images (224x224 or 299x299 depending on the model)
- Normalize pixel intensities
- Apply data augmentation to training set (rotation, flipping and zoom)

3) Step 3: Dataset Partitioning

- Split datasets into 80% training, 10% validation and 10% test

4) Step 4: Model Selection

- Select pre-trained models (VGG16, ResNet50, InceptionV3, EfficientNet)
- Customize the top layer for five-class classification (Softmax activation)

5) Step 5: Training and Validation

- Fine-tune pre-trained models on training data
- Monitor validation loss and accuracy to evaluate convergence

6) Step 6: Testing and Evaluation

- Test model performance on unseen test set
- Evaluate using metrics: accuracy and categorical cross-entropy loss

7) Step 7: Comparative Analysis

- Analyze model performance across datasets
- Interpret results with a focus on convergence, generalization, and architecture-specific behaviour

B. Datasets

1) Dataset 1: Contains 3,662 preprocessed grayscale fundus images filtered using Gaussian noise reduction, resized to 224x224 pixels, categorized into five DR stages.

2) Dataset 2: Comprises 35,126 fundus images from Kaggle [x], with varied resolution resized as per model input requirements.

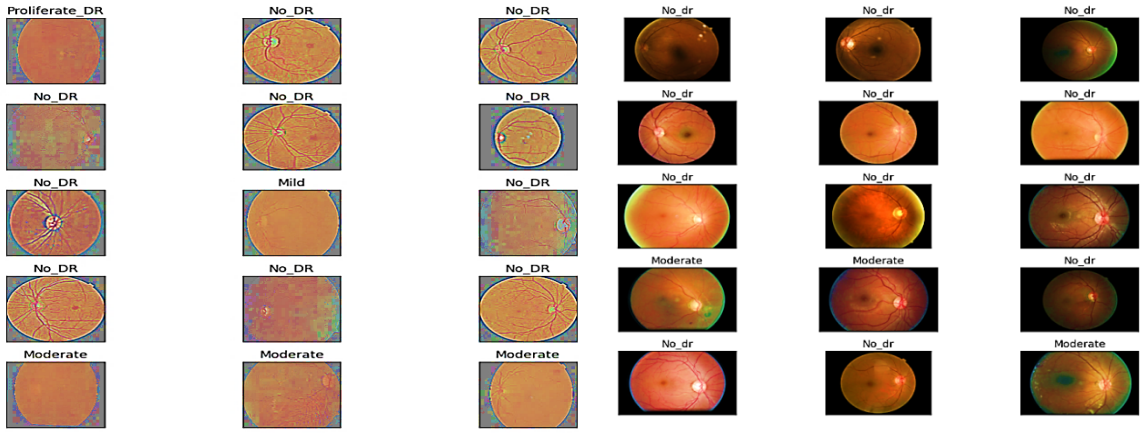


Fig 1: The datasets : (a) Dataset 1 and (b) Dataset 2

The following displays the distribution of sample photos throughout the Kaggle dataset's various classes:

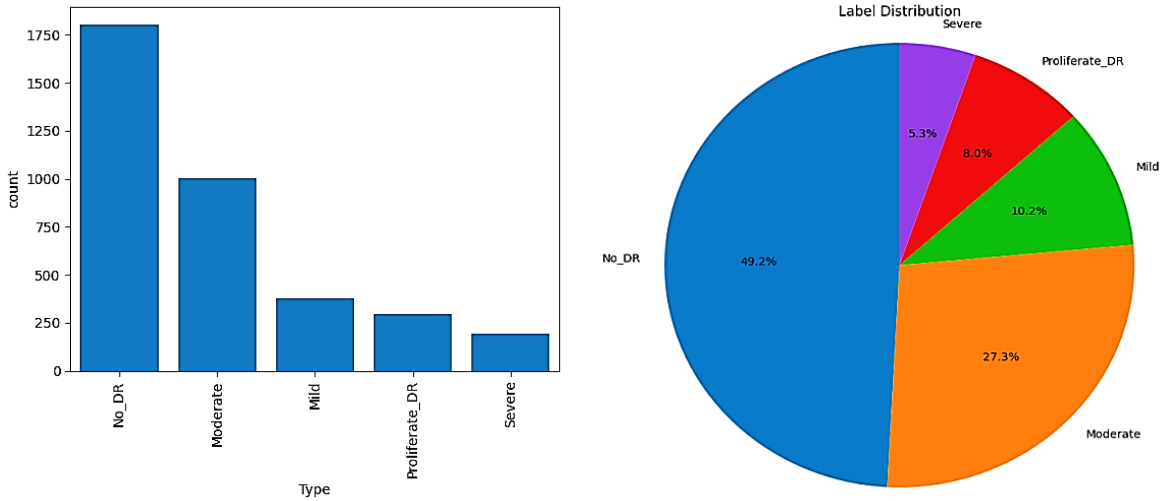


Fig 2: Distribution of different classes in table and pie chart for Dataset 1

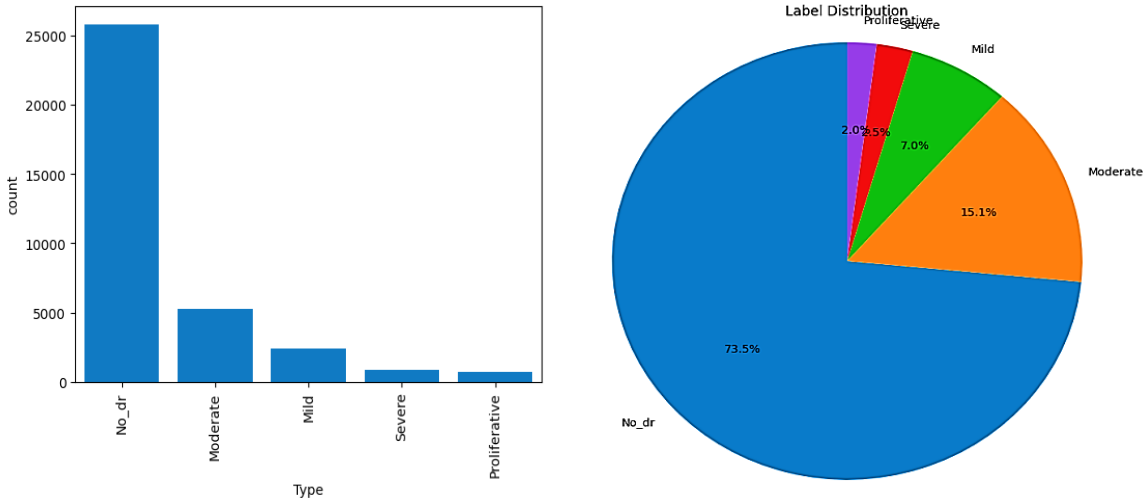


Fig 3: Distribution of different classes in table and pie chart for Dataset 2

C. Analysis of the DR Progression Curve:

1) Initial stage (No DR to Mild):

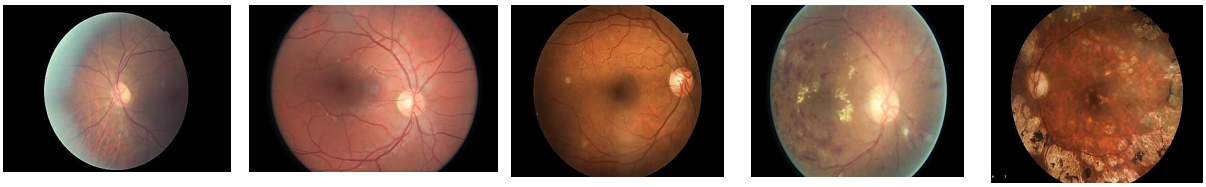
- Few abnormalities.
- Microaneurysms start forming.
- Changes are subtle, often not affecting vision.

2) Intermediate stage (Moderate):

- Retinal damage becomes more visible.
- More frequent hemorrhages and hard exudates.
- Vision may start being affected.

3) Advanced stage (Severe to PDR):

- Significant retinal ischemia.
- Neovascularization and fibrous tissue proliferation.
- High risk of vision loss or blindness.



(a) No DR (b) Mild (c) Moderate (d) Severe (e) PDR

Fig 4: Different stages of DR

D. Preprocessing

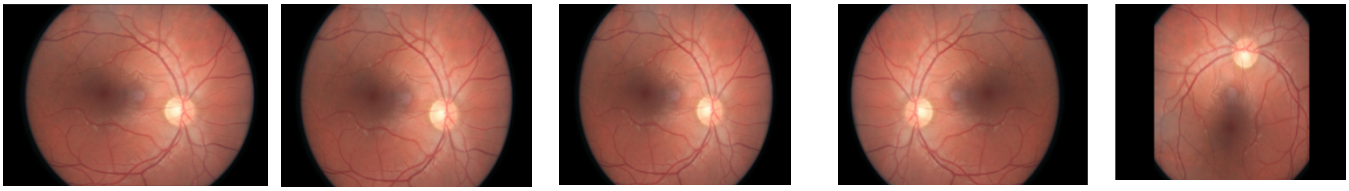
Preprocessing improves input quality and ensures compatibility with deep learning models:

1) Grayscale conversion (Dataset 1)

2) Image resizing: 224x224 (VGG16, ResNet50, EfficientNet), 299x299 (InceptionV3)

3) Normalization: Pixel values scaled to [0,1]

4) Augmentation: Applied only on training data to increase variance and mitigate overfitting



(a) Resize(Aspect Ratio) (b) Crop (c) Normalize (d) Augment (e) Rotate

Fig 5: Stages of Preprocessing

E. Model Architectures

Four CNN models were chosen for comparative evaluation:

TABLE I
MODEL ARCHITECTURES

Model	Parameters	Input size	Activation	Output	Highlights
VGG16	~138MThe following	224x224	ReLU	Softmax	Simple layered structure,

	displays the distribution of sample photos throughout the Kaggle dataset's various classes.				high parameter count
ResNet50	~25.6M	224x224	ReLU	Softmax	Residual connections enable deeper networks
InceptionV3	~23.8M,	299x299	ReLU	Softmax	Multi-scale convolutional capture diverse features
EfficientNet	~5.3M	224x224	Swish	Softmax	Lightweight model with compound scaling

F. Training Configuration

1) Batch Size: 32
2) Epochs: 5 (to monitor early convergence trends)
3) Loss Function: Categorical Cross entropy
4) Optimizer: Adam (learning rate = 0.001)
5) Metrics: Accuracy and loss on both validation and test sets

This structured methodology enables reproducibility and facilitates an objective comparison of the selected architectures under consistent experimental conditions:

TABLE 2
TRAINING CONFIGURATION

Model	Parameters	Architectural Highlights
VGG16	138M	Deep stacked convolutional layers, simple design
ResNet50	25.6M	Residual connections, identity mappings
InceptionV3	23.8M	Multi-scale convolutions, factorized filters
EfficientNet	5.3M	Compound scaling of depth/width/resolution, MBConv blocks

III. RESULTS AND DISCUSSION

The comparative performance of the four deep learning models is analyzed over five training epochs across two datasets. The results are visualized and interpreted using tabular summaries and plotted trends to provide deeper insights.

1) Model Accuracy Over Epochs (Tabular Summary)

TABLE 3
MODEL ACCURACY

Epoch	ResNet50 (Acc D1 / D2)	VGG16 (Acc D1 / D2)	InceptionV3 (Acc D1 / D2)	EfficientNet (Acc D1 / D2)
0	0.498 / 0.736	0.560 / 0.731	0.600 / 0.734	0.442 / 0.722
1	0.526 / 0.737	0.637 / 0.734	0.664 / 0.734	0.498 / 0.731
2	0.547 / 0.737	0.652 / 0.734	0.673 / 0.734	0.498 / 0.731
3	0.558 / 0.737	0.652 / 0.734	0.656 / 0.734	0.498 / 0.731
4	0.572 / 0.737	0.662 / 0.734	0.688 / 0.734	0.495 / 0.731

Key Insight: InceptionV3 consistently performed the best on Dataset 2, while ResNet50 demonstrated gradual learning on Dataset 1.

2) Model Loss Over Epochs (Tabular Summary)

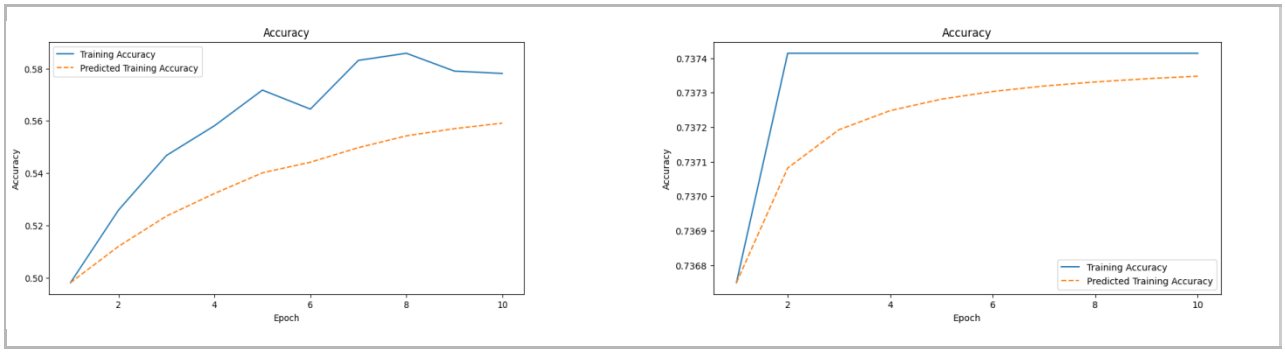
TABLE 4
MODEL LOSS

Epoch	ResNet50 (Loss D1 / D2)	VGG16 (Loss D1 / D2)	InceptionV3 (Loss D1 / D2)	EfficientNet (Loss D1 / D2)
0	1.284 / 0.873	1.174 / 0.913	1.221 / 0.845	1.958 / 1.136
1	1.196 / 0.867	0.976 / 0.890	0.927 / 0.845	1.459 / 0.955
2	1.125 / 0.866	0.945 / 0.885	0.903 / 0.845	1.424 / 0.904
3	1.119 / 0.865	0.941 / 0.882	0.931 / 0.843	1.415 / 0.886
4	1.104 / 0.865	0.920 / 0.879	0.870 / 0.843	1.414 / 0.879

Key Insight: InceptionV3 and ResNet50 show rapid convergence with continuous reduction in loss, whereas EfficientNet struggles to optimize effectively on Dataset 1.

3) Accuracy Trend Visualization

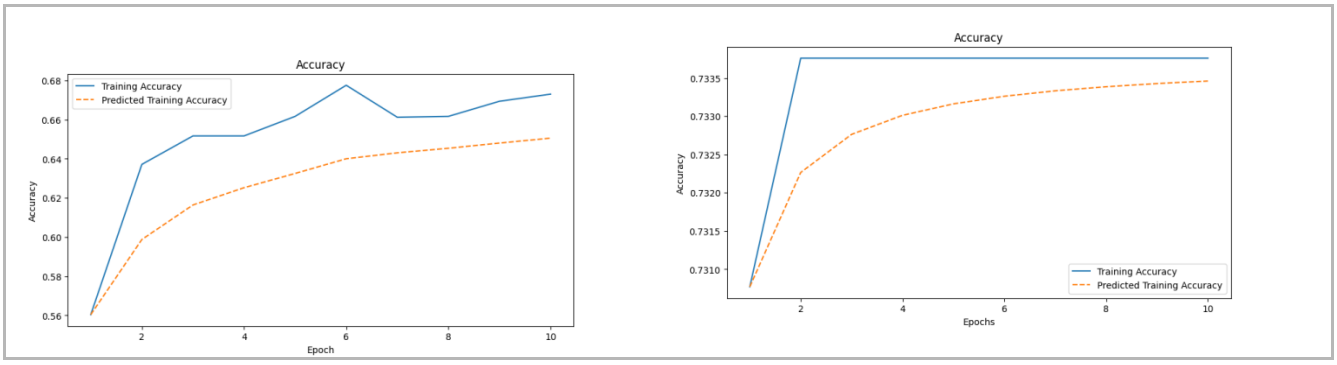
- ResNet50: Linear improvement on Dataset 1 suggests gradual feature learning. Stability on Dataset 2 from epoch 1 implies fast convergence.



(a) (b)

Fig 6: Accuracy of RestNet50 on (a) Dataset 1 and on (b) Dataset 2

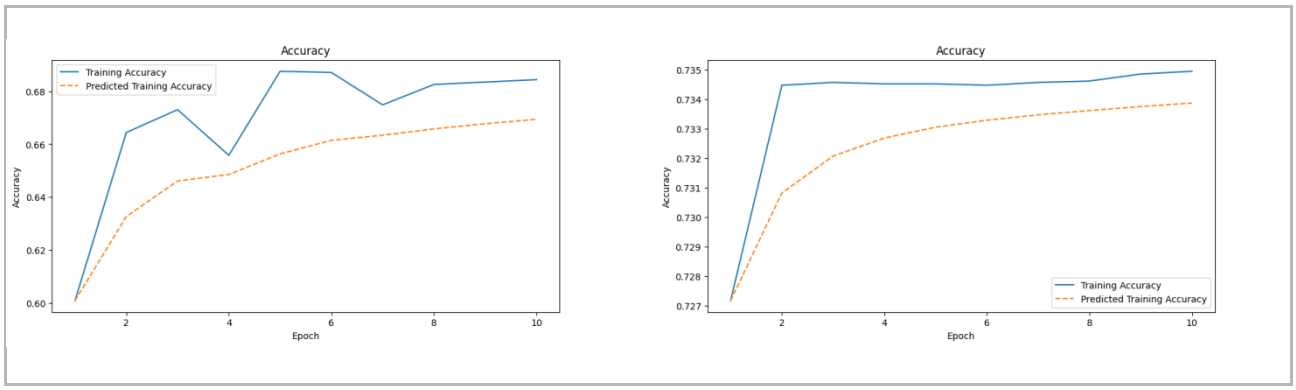
- VGG16: Steep learning curve initially on Dataset 1, followed by saturation. Consistent but slightly underwhelming on Dataset 2.



(a) (b)

Fig 7: Accuracy of VGG16 on (a) Dataset 1 and on (b) Dataset 2

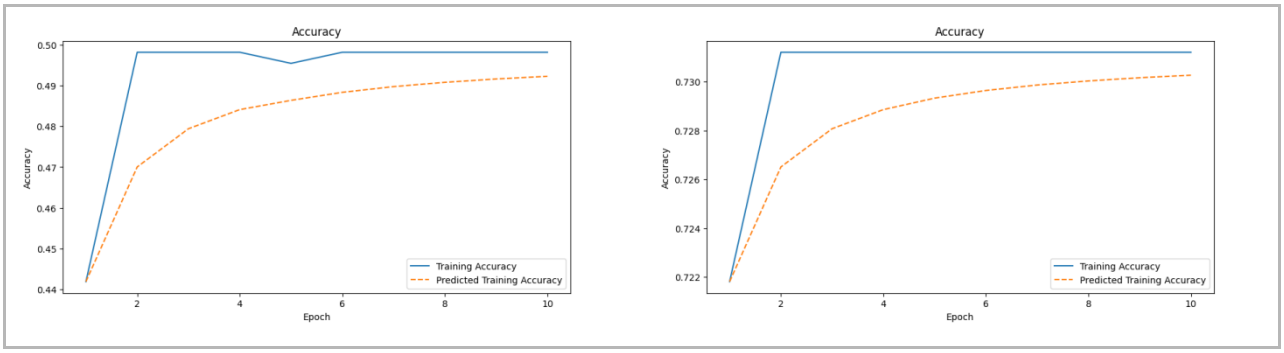
- InceptionV3: Best performer on both datasets. Smooth and steady rise in accuracy with minimal variance.



(a) (b)

Fig 8: Accuracy of InceptionV3 on (a) Dataset 1 and on (b) Dataset 2

- EfficientNet: Lagging across datasets. Marginal improvements across epochs, reflecting limited capacity for complex patterns.



(a) (b)

Fig 9: Accuracy of EfficientNet on (a) Dataset 1 and on (b) Dataset 2

4) Graphical Summary (Descriptive)

Imagine a grouped bar chart showing model accuracies side by side for each epoch. In such a chart:

- InceptionV3 bars remain the tallest across Dataset 2.
- ResNet50 bars gradually rise for Dataset 1.
- VGG16 bars start tall but plateau.

EfficientNet bars remain the shortest, indicating poor generalization.

5) Comparative Discussion by Architecture

- ResNet50: Residual connections enable deeper feature abstraction, which is particularly beneficial for low-contrast and noisy images in Dataset 1.
- InceptionV3: Its inception modules capture multi-scale features effectively, proving ideal for diverse and large-scale data as seen in Dataset 2.
- VGG16: While capable of rapid early learning, it lacks architectural enhancements like residual or attention mechanisms, leading to early saturation.
- EfficientNet: Despite its efficient design, the trade-off in representational power limits its accuracy, especially on smaller datasets.

6) Dataset Dependency and Generalization

Dataset characteristics play a pivotal role in model performance:

- Dataset 1 (smaller, grayscale, pre-filtered) challenges lightweight models like EfficientNet.

- Dataset 2 (larger, color, diverse) favors complex models like InceptionV3, which generalize better across varied feature types.

This critical analysis reveals that model architecture significantly affects performance in DR classification.
ResNet50: Best for smaller, less diverse datasets
InceptionV3: Ideal for large, complex datasets
VGG16: Rapid learner but prone to early saturation
EfficientNet: Computationally efficient but lacks performance depth
Selecting a deep learning model for medical imaging tasks should consider not just accuracy but also convergence behavior, dataset characteristics, and architectural complexity.

IV. CONCLUSION

As stated in the introduction section, the focus is to identify the most suitable model based on classification accuracy, convergence behavior and generalization ability, thereby contributing to evidence-based model selection in automated DR detection, we found InceptionV3 and VGG16 did particularly well on the smaller D1dataset because they could latch onto more subtle features quickly and EfficientNet underperformed on D1 in these epochs—this could be due to:

- Architecture complexity requires more examples to fully leverage its capacity.
- Hyperparameters (learning rate, augmentations) not yet optimized for the smaller dataset.

On the D2 dataset, all models stabilize around 0.73, suggesting that regardless of architecture, the dataset's inherent difficulty sets a performance ceiling at those early epochs.
However, the vast scope of future advancement of this work can be recognized by the use of ReduceLROnPlateau or CosineAnnealingLR for adaptive learning and helps models avoid getting stuck in local minima. As DR datasets often have skewed class distribution, use of class weighting or focal loss can be applied to boost minority class detection.
As of recent studies up to 2024, EfficientNet generally outperforms VGG16, ResNet50, and InceptionV3 in the task of diabetic retinopathy (DR) detection—both in terms of accuracy and efficiency. Here's a summary comparison based on recent literature and benchmarks on datasets like APTOS, Messidor, and EyePACS:

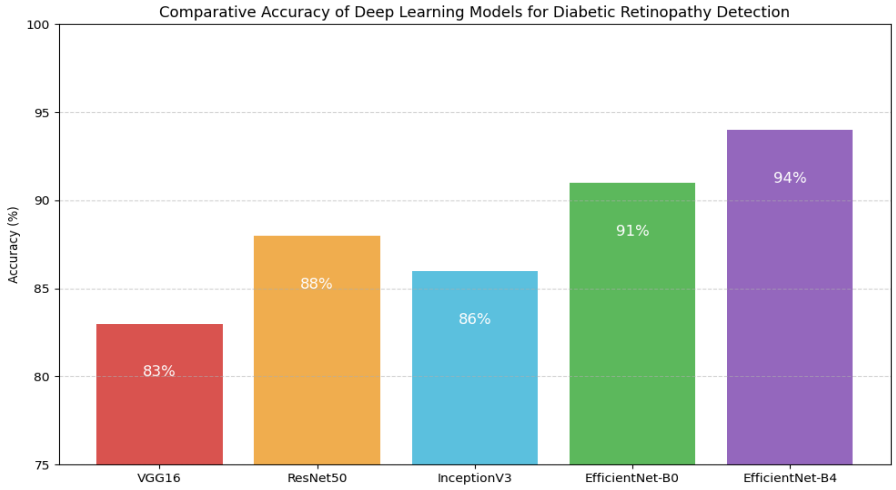
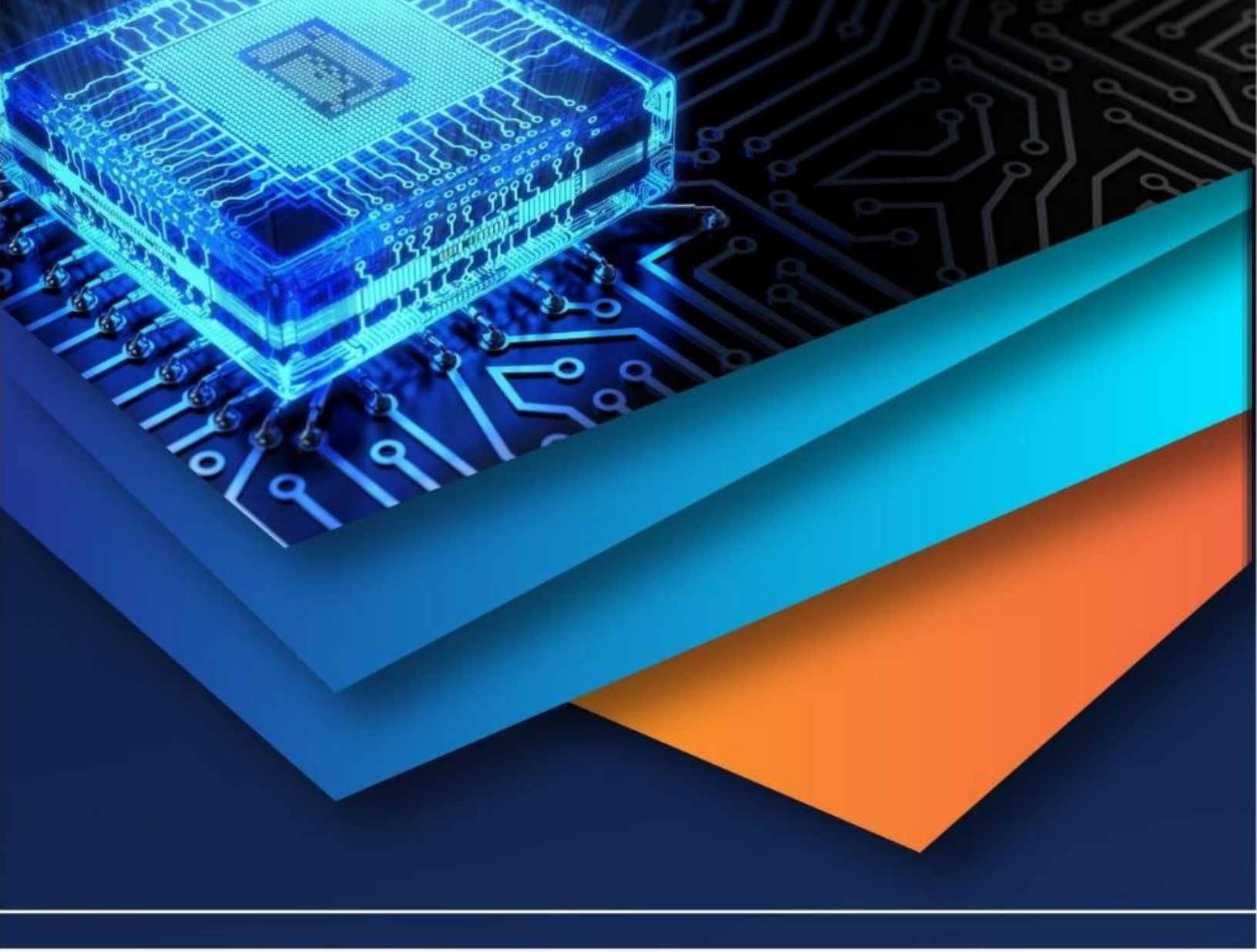


Fig10: Comparative Accuracy of DL Models

REFERENCES

[1] Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition.
[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition.
[3] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., &Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision.
[4] R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.
[5] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.
[6] Kaggle Diabetic Retinopathy Detection Dataset 1 https://www.kaggle.com/datasets/sovitrath/diabetic-retinopathy-224x224-gaussian-filtered
[7] Kaggle Diabetic Retinopathy Detection Dataset 2 https://www.kaggle.com/datasets/amanneo/diabetic-retinopathy-resized-arranged



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY