KFUPM
Journal of Undergraduate Research International

Diamond Open Access

Article — *Computer Engineering*

Identification of Diabetic Retinopathy Using Deep Learning and Ensemble Model Approach

Nusrat Mahee and Md. Sabbir Ejaz*

Cite https://doi.org/10.64589/juri/214968

Submitted: June 04, 2025 **Revised:** November 06, 2025 **Accepted:** December 01, 2025

ABSTRACT

One of the most prevalent complications of diabetes and a leading cause of preventable blindness worldwide is diabetic retinopathy (DR). Although several approaches have been employed in the existing literature for DR detection and classification, several critical limitations remain owing to the use of small or single-center datasets and a lack of cross-model benchmarking. These have resulted in reduced generalizability, poor sensitivity, and unquantified performance advantages. This study examines the history of DR detection methods and transition of conventional image-processing algorithms to state-of-the-art deep learning and ensemble models. This study employed the MESSIDOR dataset, which includes retinal fundus images with labels indicating DR severity. Images were enhanced and preprocessed, and deep learning-based features were extracted using several pretrained architectures (e.g., ResNet50, InceptionV3, DenseNet variants, VGG16, Xception) and a custom deep neural network (DNN). Additionally, ensemble methods were developed by integrating CNN-based feature extractors with machine learning classifiers, such as random forest (RF), support vector machine (SVM), XGBoost, and LightGBM, to improve the generalization and classification accuracy. Experimental results revealed that Xception achieved the highest performance, with a testing accuracy of 95%. Furthermore, EfficientNetB3 integrated with XGBoost achieved the highest testing accuracy (86.83%). This study highlights the possibility of deploying deep learning and ensemble AI in DR screening devices, particularly in telemedicine. Such an integration can significantly improve early detection rates, reduce the burden on healthcare providers, and make diabetic eye care more accessible in underserved regions.

Keywords: diabetic retinopathy, ensemble method, deep learning, machine learning, MESSIDOR, telemedicine

1. INTRODUCTION

DR is a major cause of blindness, resulting from long-term elevated blood sugar levels that damage the retinal blood vessels. DR is divided into two major categories; one is proliferative DR (PDR) and non-proliferative DR (NPDR). In diabetes patients, NPDR affects the retinal blood vessels. The condition is further divided into mild, moderate, and severe and NPDR is the initial phase of the condition. These stages are very crucial to detect early enough and apply medical intervention before the disease advances to other more serious stages and loss of vision[1]. Our study leverages methodological and experimental advancements to improve DR detection. Figure 1 illustrates retinal conditions during different stages of DR.

2. RELATED WORK

There are many algorithms, strategies, and methodologies which have been used to detect and classify DR. Great progress has been achieved in applying machine and deep learning methods, but there are still a number of major limitations. Most existing models were developed using small or single-center datasets (e.g., DIARETDB1 and EyePACS subsets), which often lack sufficient representation across various severity grades of DR. This limitation leads to inefficient sensitivity to detect mild or non-proliferative cases and diminishes external validity in other populations and imaging equipment. Other models are based on individual CNN architectures or classical machine learning algorithms and not on cross-model benchmarking so the performance benefits remained undetermined. Moreover, the ensemble and hybrid methods are either under-explored or not optimized or computational inefficiencies tend to impede the ability to apply them in real-time. Table 1 (in the supplementary section S1) presents a summary of some past work and their contributions.

In this work, to mitigate these drawbacks of previous studies, we diversify the dataset and analyze the state-of-the-art models in a comparative manner, as well as optimize hybrid/ensemble methods to achieve better generalizability and computational efficiency. In this direction, we apply the publicly available MESSIDOR database and expand it to 12,000 images through data augmentation that improves model generalization when applied to various imaging conditions. After that, we perform a comprehensive comparative analysis of seven deep learning models (ResNet50, InceptionV3, DenseNet121, DenseNet169, VGG16, Xception, and our custom DNN) and four ensemble

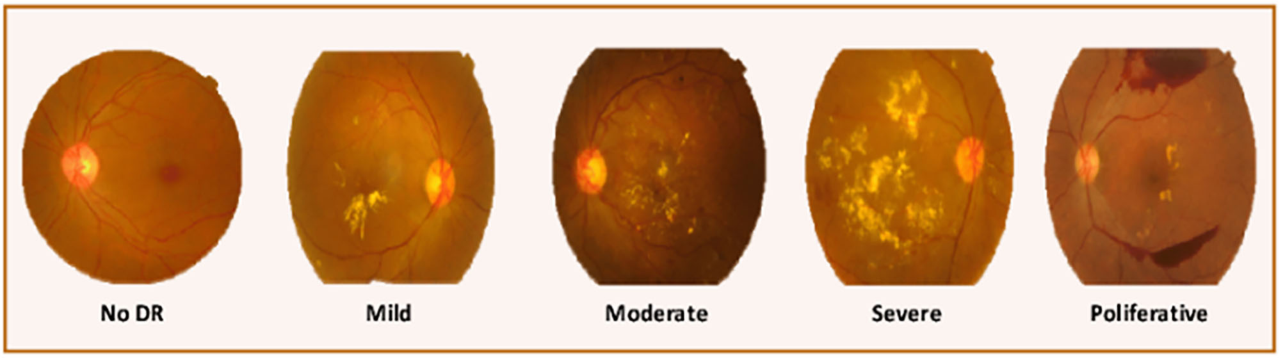

Figure 1. DR stages: (a) Normal retina; (b) mild DR; (c) moderate DR; (d) severe DR; and (e) proliferative DR[2]

models (EfficientNetB3 + XGBoost, DenseNet121 + LightGBM, ResNet50 + SVM, and ResNet50 + RF) to determine a definite hierarchy of the performance. Lastly, the hybrid and ensemble models are further optimized by using CNN-based feature extraction and advanced machine learning classifiers. Lightweight models like Xception and EfficientNetB3 are used, which are computationally efficient and clinically applicable.

3. METHODOLOGY

This study implements a systematic approach to DR detection. It employs both deep learning and ensemble models. The approach emphasizes dataset preparation, preprocessing, model selection, and training pipelines. The overall workflow of the applied methodology is illustrated in Figure 2.

3.1. Dataset Preparation. In this study, the publicly available MESSIDOR[19] dataset was used and it contained fully anonymized retinal images acquired with the ethical consent of the original providers. None of the patient information can be identified, and the data can be used in academic research. The data set has retinal fundus pictures with Grade 0 (546 pictures), Grade 1 (153 pictures), Grade 2 (247 pictures) and Grade 3 (254 pictures). We used data augmentation to increase the size of the data to 12,000 images. Augmentation techniques comprised random flips, rotations (up to 30°), color jitter, resizing to 224 × 224 pixels and affine transformations. The augmented images were saved in structured, class-wise folders while preserving the original class proportions to enhance model robustness. Table 2. summarizes the number of original and augmented images per class.

In the second step of model training, we implemented real-time preprocessing using the ImageDataGenerator by Keras. New images were not saved under this technique, but transformations were made dynamically when loading a batch. They consisted of shear transformations, zooming, horizontal flipping, rotation (up to 20°) and changes in brightness (between 0.8 and 1.2°), width and height changes. Moreover, a preprocessing input function was also used to normalize the input images according to the model architecture that is being trained.

Table 1. *Grade-wise image counts before and after augmentation*

DR grade	Original count	Augmented count	Final count
Grade 0	546	4914	5460
Grade 1	153	1377	1530
Grade 2	247	2223	2470
Grade 3	254	2286	2540
Total	**1200**	**10800**	**12000**

3.2. Overview and Rationale of Implemented Models. Our study utilized seven deep learning models and four ensemble models. The deep learning models include ResNet50, InceptionV3, DenseNet121, DenseNet169, Xception, VGG16, and DNN. These models were selected because they represent diverse and well-established CNN architectures. ResNet and DenseNet address gradient flow through residual and dense connections, InceptionV3 and Xception provide efficient multi-scale and depthwise-separable feature extraction, and VGG16 offers a strong convolutional baseline. A custom DNN is included as a lightweight benchmark. Together, they allow a comprehensive evaluation of different feature-learning mechanisms for DR classification.

The ensemble models include EfficientNetB3 + XGBoost, ResNet50 + RF, ResNet50 + SVM, and DenseNet121 + LightGBM. The primary variations among the ensemble models lie in the choice of deep feature extractor and machine learning classifier. The EfficientNetB3 + XGBoost model combines a compact yet efficient CNN with gradient boosting to achieve high accuracy and interpretability. The ResNet50 + RF model combines the residual training of ResNet and the resiliency of decision-tree ensembles to enhance generalization. The ResNet50 + SVM model integrates deep residual features with a linear kernel SVM, which is particularly suitable for high-dimensional feature spaces. Finally, the DenseNet121 + LightGBM model leverages the densely connected CNN layers to generate rich features and uses them in a LightGBM model that is constrained to multiclass classification based on the log loss. These architectural variations influence how each model captures fine-grained features and balances the bias–variance trade-off, thereby affecting classification performance and error behavior. An overview of the implemented models is included in the supplementary section S2.

3.2.1. Deep learning model configuration. In this study, we applied transfer learning employing the pretrained CNN

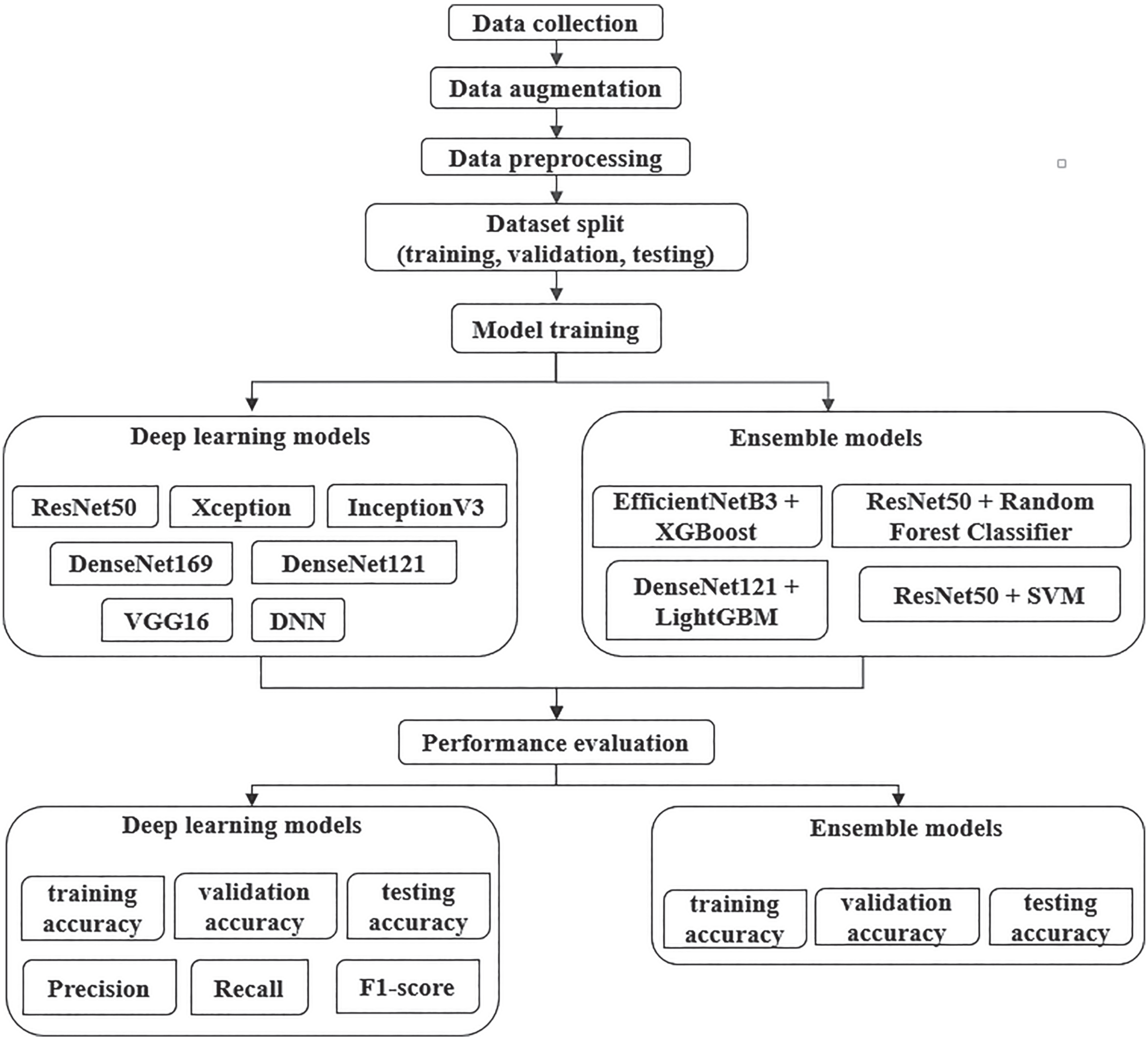

Figure 2. Schematic representation of the workflow

architectures of the implemented models. These models, originally trained on ImageNet, provide strong feature-extraction capabilities. We tailored the models for our task by selectively unfreezing 4–20 of the final layers depending on the model. These layers were fine-tuned to learn high-level retinal features, whereas low-level features from pretraining were retained.

We replaced the original classifier with a fully connected custom head, starting with a global average pooling layer to reduce the dimensions, followed by dense layers (1024 and 512 units) with rectified linear unit activation. We included dropout (rate of 0.3) and L2 regularization (0.001) to prevent overfitting. The final layer used softmax activation for multiclass classification. Training was performed using the Adam optimizer (learning rate = 0.0001) and categorical cross-entropy loss.

3.2.2. Ensemble model configuration. In our study, all four ensemble models follow a similar ensemble pipeline that combines deep learning-based feature extraction with classical machine learning classifiers. A pretrained CNN, EfficientNetB3, ResNet50, or DenseNet121, is loaded without including its final layers and treated as a feature extractor. These extracted features are then fed into a classic machine learning model (e.g., XGBoost, SVM, RF, or LightGBM), which is trained as a multiclass classifier. After training, all models are assessed using accuracy measures, and classification reports, which provide quantitative and visual measures of the predictive capacity of each model at each DR severity level.

3.2.3. Hardware and software environment. All experiments were conducted in a Windows-based system using Jupyter Notebook (Anaconda Navigator). The models were implemented in Python. The same hardware and software environment was used across all models to ensure consistency, fairness, and reproducibility.

4. RESULTS AND DISCUSSION

This section summarizes the results obtained from implementing the deep learning and ensemble models on the MESSIDOR

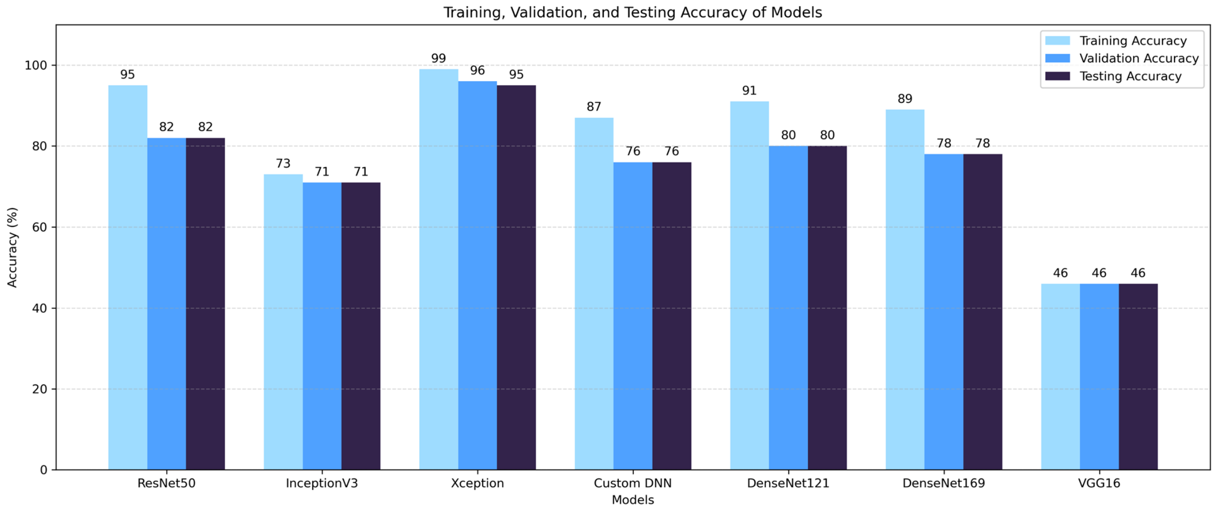

Figure 3. Bar chart of training, validation, and testing accuracies for deep learning models

dataset. The analysis is organized into four parts: Section 4.1 presents the results of the deep learning models, and Section 4.2 discusses the performance of the ensemble models. Section 4.3 provides a comparative analysis of existing studies, and Section 4.4 offers a detailed discussion of the key findings.

4.1. Deep Learning Model Results. In this study, we evaluated seven deep learning models for DR detection. The training, validation, and testing accuracies are presented in Table 3 (provided in the supplementary section S3), and Figure 3.

Xception demonstrated the highest performance, with training, validation, and testing accuracies of 99%, 96%, and 95%, respectively, showing excellent generalization. ResNet50 trained with 95% accuracy but only 82% on validation and testing, indicating a marginal overfitting. Smaller overfitting (91% training, 80% testing, 89% training, and 78% testing) in DenseNet121 and DenseNet169 also indicated the successful extraction of features, as well as the memorization of training features. The custom DNN scored 87% training and 76% testing, indicating moderate overfitting, as the model fit some non-generalizable patterns. InceptionV3 scored worse in validation and testing (71%), likely owing to underfitting or architectural incompatibility. VGG16 achieved the lowest accuracy (46%) due to its shallow, uniform architecture and lack of residual connections, which limit gradient flow and the ability to capture fine-grained retinal features. Consequently, it struggles to distinguish subtle lesions across DR severity levels. Table 4 outlines the accuracy, recall, and F1-scores of the models for the four DR grades in the MESSIDOR dataset.

A comparative analysis of the models on the MESSIDOR dataset reveals that Xception was the most successful, with the highest precision, recall, and F1-scores across all DR grades. This excellent performance can be explained by its depthwise separable convolutions and residual connections that enable highly efficient feature extraction by decoupling channels and spatial learning. Consequently, Xception can capture fine retinal lesions while remaining computationally efficient.

ResNet50 and DenseNet169 also exhibited robust performance owing to their deep residual and dense connections, which are more effective in learning hierarchical lesion patterns. Conversely, although both InceptionV3 and VGG16 are stable and interpretable, they achieved relatively lower performance because they are both shallow and less optimized for complex feature hierarchies. Custom DNN performed poorly due to pure dense structures that do not possess the spatial feature extraction. Generally, models with an effective convolutional design and feature reusability were highly effective in differentiating DR severity levels.

4.2. Ensemble Model Results. This section summarizes the performances of the four ensemble models. A comprehensive

Table 2. *Class-wise precision, recall, and F1-scores across DR grades*

Model	Precision across grades				Recall across grades				F1-score across grades			
	0	1	2	3	0	1	2	3	0	1	2	3
Xception	0.94	0.93	0.95	0.99	0.98	0.89	0.92	0.96	0.96	0.91	0.93	0.97
ResNet50	0.87	0.60	0.77	0.92	0.85	0.76	0.74	0.87	0.86	0.67	0.75	0.89
DenseNet121	0.76	0.61	0.76	0.90	0.91	0.56	0.61	0.73	0.83	0.58	0.68	0.81
DenseNet169	0.80	0.68	0.67	0.94	0.90	0.45	0.77	0.75	0.85	0.54	0.72	0.84
Custom DNN	0.46	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.63	0.00	0.00	0.00
InceptionV3	0.70	0.68	0.64	0.84	0.90	0.39	0.51	0.70	0.79	0.49	0.57	0.76
VGG16	0.77	0.61	0.66	0.94	0.91	0.49	0.64	0.70	0.83	0.54	0.65	0.80

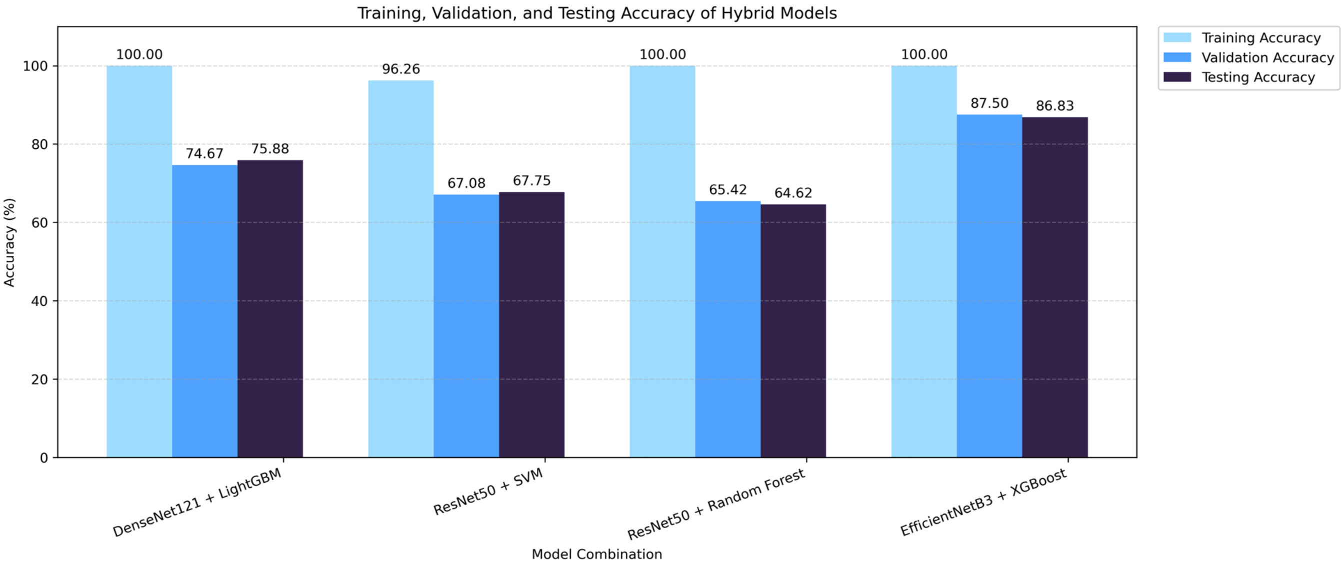

Figure 4. Bar chart of training, validation, and testing accuracies for ensemble models

summary of the experimental results is presented in Table 5 (provided in supplementary section S4) and illustrated in Figure 4.

EfficientNetB3 + XGBoost achieved training, validation, and test accuracies of 100%, 87.50%, and 86.83%, respectively. The gap between training and validation/testing indicates overfitting. The ensemble memorized the training patterns but was still adequately generalized to unseen data. The highest test accuracy can be attributed to the strong feature extraction capability of EfficientNetB3. The model is able to capture the finer retinal details. The enhanced robustness and classification capability were caused by XGBoost capacity to capture the complicated relationship between these features. The DenseNet121 + LightGBM also had 100% training and lower validation (74.67%) and testing (75.88%) performances due to overfitting caused by memorization of training features. ResNet50 + SVM and ResNet50 + RF had good training accuracy (96–100%) and poor test accuracy (67.75% and 64.62%, respectively), which were characteristics of overfitting and ineffective feature integration.

4.3. Comparison with Existing Works.

Compared with previous studies, our study on the MESSIDOR dataset demonstrates competitive and, in several cases, superior performance. Earlier research using MESSIDOR achieved notable results, such as VGG16 (94.57%), DenseNet121 (79.75%)[8], IoTDL-DRD (99.08%)[9], and ResNetGB (98.88%)[10], often relying on complex ensemble models or optimization-based techniques. By contrast, our implementation of the Xception model achieved a test accuracy of 95%, demonstrating strong performance without heavy reliance on metaheuristic algorithms. In addition, our ensemble model, EfficientNetB3 + XGBoost, outperformed several ensemble models from previous studies, such as the ensemble of ResNet50, InceptionV3, Xception, DenseNet121, and DenseNet169, which reached only 80.08%[1], and the VGG16 + XGBoost model, which reached 79.5%[11] by achieving 86.83% test accuracy. Although some studies on other datasets, such as EyePACS and DIARETDB1, have reported better results (ResNetGB: 99.73%, Modified U-Net: 98.68%), these datasets are typically larger or more diverse. Despite the moderate size and limited variability of MESSIDOR, our models (e.g., DenseNet121: 80%; ResNet50: 82%) showed strong learning capabilities. This highlights that with effective data augmentation, model fine-tuning, and thoughtful ensemble combinations, a high performance in DR detection is achievable, even on more constrained medical imaging datasets.

4.4. Discussion of the Findings.

Early training revealed diverse behaviors across architectures. DenseNet121 + LightGBM and EfficientNetB3 + XGBoost both showed 100% training accuracy with lower validation/testing (74.67%–87.5%), indicating overfitting but reasonable generalization. VGG16 underfit with 46% accuracy, while InceptionV3 and ResNet50 showed moderate training (>70%) but reduced validation/testing performance. These results suggest that although advanced deep learning architectures effectively capture complex retinal features, the performance of the ensemble models can be further improved by combining the strengths of models. Nevertheless, the findings also suggest that the success of ensemble-based methods is highly dependent on the combination of the best models and the tuning of hyperparameters to make the ensemble approach robust and generalizable.

This study has some limitations. It uses only the MESSIDOR dataset, which may limit how well the results apply to other populations or imaging conditions. Only a few ensemble combinations were tested, and important validation measures like AUC and cross-validation were not fully included. Future research should include larger and more diverse datasets and explore additional ensemble strategies. It should also focus on optimizing model architectures and incorporating supplementary validation metrics to improve robustness and diagnostic accuracy.

5. CONCLUSIONS

This study evaluated deep learning and ensemble models for DR detection using the MESSIDOR dataset. Xception achieved the highest test accuracy (95%), followed by ResNet50 (82%) and DenseNet121 (80%), confirming the strong feature-extraction capability of advanced CNN architectures. The EfficientNetB3

+ XGBoost ensemble reached a notable test accuracy of 86.83%, outperforming some standalone models and demonstrating the value of well-designed ensemble strategies. However, not all ensembles were effective; combinations such as ResNet50 + SVM and ResNet50 + RF performed poorly, underscoring the importance of selecting compatible feature extractors and classifiers. Clinically, the findings have important implications. High-performing models such as Xception and EfficientNetB3 + XGBoost could support early detection of DR, assist in telemedicine-based screening programs, and improve real-world clinical workflows by reducing the burden on ophthalmologists and expanding diagnostic services to underserved populations. Incorporating these AI models into routine practice has the potential to enhance patient outcomes and facilitate large-scale DR screening. Overall, the results align with previous findings and highlight the need for careful architectural and hyperparameter choices to maximize DR classification performance.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Nusrat Mahee – *Department of Information and Communication Engineering, Noakhali Science and Technology University, Bangladesh;* 0009-0004-6103-5175
Email: mahee1115@student.nstu.edu.bd

Corresponding Author

Md. Sabbir Ejaz – *Department of Information and Communication Engineering, Noakhali Science and Technology University, Bangladesh;* 0009-0001-1789-2258
Email: sabbirejaz.ice@nstu.edu.bd

AUTHOR CONTRIBUTION STATEMENT

Nusrat Mahee: Collection and/or assembly of data, data analysis and interpretation, writing the article. **Md. Sabbir Ejaz:** Research concept and design, critical revision of the article, final approval of the article.

ACKNOWLEDGMENTS

I would like to thank my supervisor, Md. Sabbir Ejaz, who has helped me with this research in terms of his constant guidance, encouragement, and helpful ideas. I also acknowledge the Department of Information and Communication Engineering (ICE), Noakhali Science and Technology University (NSTU) for providing the resources and conducive environment that made this work possible.

FUNDING

This research was financially supported by the NSTU Research Cell, Noakhali Science and Technology University (NSTU), which is gratefully acknowledged.

CONFLICTS OF INTEREST

The author states that there are no financial, personal or professional conflict of interest and denies any competing interests with regards to the work.

REFERENCES

(1) Qummar, S., *et al.* A deep learning ensemble approach for DR detection. *IEEE Access.* **7**, 150530–150539 (2019).

(2) Bashir, I. *et al.* RDS-DR: An Improved Deep Learning Model for Classifying Severity Levels of DR. *Diagnostics,* **13**, 3116 (2023).

(3) Deepa, V., Kumar, C. S., & Cherian, T. Ensemble of multi-stage deep convolutional neural networks for automated grading of DR using image patches. *Journal of King Saud University- Computer and Information Sciences*, **34**(8), 6255–6265 (2022).

(4) P. R., & A. P. Diagnosis of DR using machine learning techniques. *ICTACT Journal on Soft Computing,* **3**(4), 563–575 (2013).

(5) S. Dutta, B. C. S. Manideep, S. M. Basha, R. D. Caytiles, & N. C. S. N. Iyengar. Classification of DR images by using deep learning models. *International Journal of Grid and Distributed Computing*, **11**(1), 89–106 (2018).

(6) Naveen, R., Sivakumar, S. A., Maruthi Shankar, B., & Keerthana Priyaa, A. DR detection using image processing. *International Journal of Engineering and Advanced Technology*, **8**(6 Special Issue), 937–941 (2019).

(7) Carrera, E. V., Gonzalez, A., & Carrera, R. Automated detection of DR using SVM. Proc. 2017 IEEE 24th International Congress on Electronics, Electrical Engineering and Computing (INTERCON), (2017).

(8) T. Swapna, D. Akhila, P. Srija, T. Shivani, and P. Srinidhi, DR Classification using Transfer Learning. *International Journal of Advanced Trends in Computer Science and Engineering,* **12**(3), (2023).

(9) T. Palaniswamy and M. Vellingiri, Internet of Things and Deep Learning Enabled DR Diagnosis Using Retinal Fundus Images. *IEEE Access*, **11**, 27590–27601 (2023).

(10) F. Saeed, M. Hussain, and H. A. Aboalsamh, Automatic DR Diagnosis Using Adaptive Fine-Tuned Convolutional Neural Network. *IEEE Access*, **9**, 41344–41359 (2021).

(11) Mohanty, C., *et al.* Using deep learning architectures for detection and classification of DR. *Sensors*, **23**(12), 5726, (2023).

(12) A. Erciyas and N. Barışçı, An Effective Method for Detecting and Classifying DR Lesions Based on Deep Learning. *Computational and Mathematical Methods in Medicine*, **2021**(1), 9928899 (2021).

(13) A. Skouta, A. Elmoufidi, S. Jai-Andaloussi, O. Ouchetto, Hemorrhage semantic segmentation in fundus images for the diagnosis of DR by using a convolutional neural network. *Journal of Big Data*, **9**(1), (2022).

(14) R. Reguant, S. Brunak, & S. Saha. Understanding inherent image features in CNN-based assessment of DR. *Scientific Reports,* **11**(1), 1–12 (2021).

(15) Ramasubramanian, B. & Prabhakar, G. An Early Screening System for the Detection of DR using Image Processing. *International Journal of Computers and Applications*, **61**, 6–10 (2013).

(16) Gao, Z. *et al.* Diagnosis of DR Using Deep Neural Networks. *IEEE Access* **7**, 3360–3370 (2019).

(17) Samanta, A., Saha, A., Satapathy, S. C., Fernandes, S. L. & Zhang, Y. D. Automated detection of DR using convolutional neural networks on a small dataset. *Pattern Recognition Letters*, **135**, 293–298 (2020).

(18) Zhao, Y. *et al.* Using Machine Learning Techniques to Develop Risk Prediction Models for the Risk of Incident DR Among Patients with Type 2 Diabetes Mellitus: A Cohort Study. *Frontiers in Endocrinology (Lausanne)*, **13**, 876559 (2022).

(19) E. Decencière, X. Zhang, G. Cazuguel, B. Laÿ, B. Cochener, *et al.*, Feedback on a publicly distributed image database: The Messidor database. *Image Analysis & Stereology*, **33**(3), 231–234 (2014).