# Diabetic Retinopathy Classification: A Cascading Model Approach

Abhirup Paul
*School of Computing*
*Dublin City University*
Dublin, Ireland
abhirup.paul2@mail.dcu.ie

Atharva Patil
*School of Computing*
*Dublin City University*
Dublin, Ireland
atharvarajkumar.patil2@mail.dcu.ie

Claudia Mazo
*School of Computing*
*Dublin City University*
Dublin, Ireland
ORCID: 0000-0003-1703-8964
claudia.mazo@dcu.ie

***Abstract*—Diabetic Retinopathy (DR) is one of the leading causes of vision deficiency and blindness among diabetic patients worldwide. It occurs due to damage to the blood vessels in the retina, particularly in the light-sensitive tissue at the back of the eye. Early detection of DR is crucial, as it helps prevent severe outcomes such as vision loss and blindness through timely diagnosis. This paper presents a cascade model approach for the classification of DR, integrating transfer learning with Convolutional Neural Networks to optimize feature extraction and classification accuracy. Two publicly available datasets were used: the APTOS2019 dataset with 3,662 images, and the Diabetic Retinopathy Resized dataset with 35,126 images. Both datasets classify DR into five stages: 0 (No DR), 1 (Mild), 2 (Moderate), 3 (Severe), and 4 (Proliferative DR). A multi-stage architecture is implemented to address class imbalance and enhance detection accuracy, especially between visually similar DR severity levels, using structured data augmentation and oversampling techniques. Pre-trained models, including ResNet50, MobileNet, DenseNet, and AlexNet, were evaluated, among which ResNet50 consistently demonstrated the best performance across classification rounds. The proposed cascade-based approach using ResNet50 showed strong performance across different DR stages. For early-stage classification, the model achieved F1-scores of 0.81 for No DR, 0.12 for Mild DR, and 0.32 for Moderate DR. When these early stages were grouped, the model achieved an F1-score of 0.97, along with 0.29 for Severe DR and 0.43 for Proliferative DR. These results demonstrate the model's strength distinguish between both both subtle and advanced DR stages. This framework can support automated screening systems and assist clinicians in early DR diagnosis.**

***Index Terms*—Diabetic Retinopathy, Deep Learning, Transfer Learning, Convolutional Neural Networks (CNN), Cascaded Architecture, Fundus Imaging.**

## I. Introduction

Diabetes Retinopathy (DR) is a prevalent medical condition that has become increasingly common worldwide, characterised by elevated amounts of glucose in the bloodstream [1]. DR appears when blood glucose damages the retinal blood vessels. Elevated blood glucose weakens retinal arteries, causing leakage and blurred vision. In later stages, new vessels rupture, leading to vision loss. DR poses a significant threat to vision, often leading to irreversible blindness if not detected early and treated effectively. Disease severity depends on different levels, with fundus screening being the most effective early detection method [2]. Currently, 537 million adults have diabetes, projected to rise to 643 million by 2030, with a 316 million people increase in healthcare cost over 15 years [3]. DR is classified into five stages based on the vascular abnormalities and their manifestations on the retina: (i) no abnormalities; (ii) mild Non-Proliferative DR (NPDR); (iii) Moderate NDPR; (iv) Severe NDPR; and (v) Proliferative DR (PDR).

DR is defined by the presence of different symptoms in the retina, and it is diagnosed using retinal fundus images. Convolutional Neural Network (CNN) models are trained using natural image datasets or other medical datasets that are applied to another new medical task; This approach is known as Transfer Learning (TL). TL is a crucial component in CNN and provides several solutions to these problems for medical imaging applications. Studies demonstrate that generic descriptors extracted from pre-trained CNN models are very effective when recognising and locating objects from natural images. Pre-trained CNN models functions as feature extractors to identify and capture essential patterns from the dataset images [4].

In automated DR image analysis, publicly available datasets play a crucial role in developing and benchmarking machine learning models. Firstly, the APTOS 2019 dataset, which consists of 3,662 retinal images captured using fundus photography under diverse imaging conditions. These images were collected from multiple clinics and rated by clinicians on a five-level severity scale from 0 No DR to 4 Proliferative DR, with 1,805 images classified as No DR, 370 as Mild DR, 999 as Moderate DR, 193 as Severe DR, and 295 as Proliferative DR. The dataset presents real-world challenges such as variations in imaging quality, illumination, and focus, making it an ideal benchmark for robust DR detection models. This dataset has been used in research to advance automated DR screening, aiding early detection and potential prevention of blindness, particularly in resource-limited settings [5]. In [2], [4], [6]–[11] the APTOS 2019 dataset was used to develop and evaluate different models for automated DR detection. In [8], a study of multilevel DR classification using advanced machine learning techniques like VGG16, VGG19, Xception, InceptionV3, and MobileNetV2. The researchers used a step-by-step approach to identify the presence of DR and then classify its severity into five categories. The study utilized the APTOS2019 dataset which was split into 80% training, 10% validation, and

10% testing. The study uses MobileNetV2 for a three-level evaluation: first, classifying DR vs. No DR; second, grouping Mild DR vs Proliferative DR; and third, further distinguishing between Mild vs Moderate and Severe vs Proliferative DR for detailed severity assessment. The SGD optimizer achieved the best performance with 96.73% accuracy, 97% precision, recall, and F-score, making it the most effective among the three. In contrast, the RMSprop optimizer had the lowest performance, with 88.56% accuracy, 88% recall, and 88% F-score, marking it as the least effective. In [6] a Modified XceptionNet, ResNet50, and InceptionV3 models are used for DR classification of the five classes. The dataset was split as 70% training, 20% validation, and 10% testing. The study shows ResNet50 performed the worst, with 74.64% accuracy, 56.52% sensitivity, and 85.71% specificity. In contrast, the modified XceptionNet achieved superior results, with AUC of 100% (No DR), 94% (Moderate), and 92% (Mild), while Severe DR and Proliferative DR scored 88% and 85%, respectively, demonstrating significantly better performance. In [2] transfer learning models, specifically AlexNet and DenseNet, were proposed for classifying DR stages. The dataset was split in an 80%, 10%, and 10% ratio for training, testing, and validation. Comparing DenseNet-169 and AlexNet for DR classification, DenseNet-169 performs better overall, especially in class 4 (F1-score: 0.69, Precision: 0.64, Recall: 0.74), making it the best-performing class. In contrast, AlexNet struggles the most with class 2 (F1-score: 0.15, Precision: 0.31, Recall: 0.10), showing the weakest performance. DenseNet-169 consistently achieves higher Precision and Recall, improving classification across mild, moderate, and severe DR stages. In [9] a deep transfer learning model evaluating AlexNet, ResNet18, SqueezeNet, GoogleNet, VGG16, and VGG19, for DR classification was proposed. VGG16, VGG19, and AlexNet achieve the highest accuracies in DR classification. VGG16 excels in Class 2 (98.1%), AlexNet leads in Class 0 (99.7%) and Class 1 (98.0%), while VGG19 performs well in Class 1 (98.6%) and Class 2 (97.6%). GoogleNet remains competitive, but SqueezeNet performs the worst, especially in Class 3 (67.8%) and Class 4 (80.9%), with the lowest overall accuracy (90.3%). ResNet18 also struggles in Class 4 (89.8%). AlexNet (97.9%) and VGG19 (97.4%) are the most effective models. In [7] a transfer learning approach with ResNet50 is used for the classification of DR into five classes. The model was evaluated based on accuracy and Cohen's Kappa score. The dataset was split in 80%, and 20% ratio for training and testing. The experimental results demonstrated that the model achieved an overall testing accuracy of 90% and a Cohen's Kappa score of 0.94. In [10] the effectiveness of ResNet50 with explainable AI for classifying DR into five severity levels was evaluated. Their model was trained on an augmented and resized dataset, achieving training accuracy of 96.1%, validation accuracy of 88.3%. Additionally, it recorded a precision of 0.855, a recall of 0.877, and an F1-score of 0.86. The Cohen Kappa score of 0.945 further indicates an agreement between the model's predictions, highlighting the potential of ResNet50 and Explainable AI in achieving high accuracy for DR detection. However, class-wise results were not provided, and the reported metrics are based on 20 epochs. In [11] transfer learning models were applied using VGG16, InceptionResNetV2, and ResNet50 to classify five classes of DR. The deep learning-based approach using the APTOS 2019 dataset achieved an accuracy of 89.50%, a precision of 79.93%, a recall of 64.42%, and AUC of 92.76%, indicating strong performance in AUC. In [4] multiple deep learning models, including ResNet-50, Inception-V3, InceptionResNet-V2, DenseNet-169, XceptionNet, and EfficientNet-B4, were utilized to classify DR into five stages. The APTOS 2019 dataset underwent image normalization, data over-sampling, and data augmentation to enhance model performance. Among these models, Inception-V3 with a GAP-based approach achieved the highest Quadratic Weighted Kappa (QWK) score of 82.0%. Looking at results, the studies demonstrate that the first focuses on accuracy and AUC, while the second highlights QWK, offering a deeper evaluation of model performance and showing that the models with broader architectures may offer more reliable agreement with true labels.

Secondly, the Diabetic Retinopathy (resized) dataset [12] used for DR detection consisted of 35,126 images from 17,563 patients. This dataset includes images from classes 0 to 4, 0 - No DR (25810 images), 1 - Mild DR (2443 images), 2 - Moderate DR (5292 images), 3 - Severe DR (873 images) and 4 - Proliferative DR (708 images). Studies such as [3], [13] have utilized this dataset to develop and evaluate DR detection models. [13] focuses on classifying DR into five classes of the dataset using Deep learning, specifically the VGGNet16 CNN architecture. To enhance feature extraction they applied statistical pre-processing techniques including mean, median, and standard deviation calculations. The model was trained on the DR (resized) dataset with 80% training and 20% testing dataset split, achieving an overall testing accuracy of 72.5%. The authors have only mentioned the VGGNet16 model and the overall results achieved using the resized dataset. [3] focuses on multiclass identification and classification of DR using transfer learning. The researchers utilized the RestNet50 model with the RMSprop optimizer to classify five DR stages. Their model was trained with preprocessing techniques such as image resizing, data augmentation, up-sampling, image flipping, and rotation. In this paper, they achieved an overall validation accuracy of 95.01% and a training accuracy of 98.97%, along with a Cohen's kappa score of 0.96 up to 12 epochs. However, they have not mentioned the results for the five different classes.

Thirdly, the Messidor dataset is a widely used dataset for diabetic DR, consisting of 1,200 retinal images from 1,200 patients. Each image is labeled into three DR severity levels (0: No DR, 1: Mild DR, 2: Severe DR), making it a multi-class classification dataset. The dataset includes images captured under diverse conditions and from different populations, offering a varied sample set for training and evaluating different models. Research works such as [1], [14]–[16] have used this dataset for developing and validating DR detection models. In [14], the APTOS 2019 dataset was used for training, while the Messidor dataset was used for evaluation. The classification task involved identifying five severity levels of DR. Several deep learning models, including EfficientNetV2B0, DenseNet121, ResNet50, XceptionNet, and InceptionNetV4, were implemented for DR classification. Various preprocessing techniques, such as image resizing, Canny, Sobel, Roberts, Scharr edge detection, and median and variance filtering, were applied to enhance image quality. Among all models, EfficientNetV2B0 achieved the highest performance, with an overall testing accuracy of 95%, F1-score of 93%, precision of 97%, recall of 86%, and an AUC of 0.98. The high AUC value demonstrates the model's ability to effectively distinguish between different DR severity levels while minimizing false positives. The authors of this paper has not mentioned the per-class results, but they have provided the results for each model used. EfficientNetV2B0 achieved the best performance, whereas DenseNet-121 had the lowest results. In [15] InceptionV3, MobileNet, and VGG16 were used for binary classification of DR using transfer learning. The Messidor and Messidor-2 datasets were combined to form the dataset, which was then divided into two sets training set and a testing set. The dataset was split into 80% for training and 10% for testing. After combining the datasets, the classes were named DR0 (DR Negative) and DR1 (DR Positive). Various preprocessing techniques, including image cropping, CLAHE (Contrast Limited Adaptive Histogram Equalization), and image resizing, were applied to enhance image quality. Model performance was evaluated using accuracy, precision, recall, and F1-score, which were computed using sklearn.metrics in Keras. Among the three CNN models, InceptionV3 achieved the highest overall accuracy of 84% along with precision, recall, and F1-scores of 84% (85% DR0, 84% DR1), 83% (86% DR0, 83% DR1), and 83% (85% DR0, 83% DR1) respectively, followed by MobileNet with 83% accuracy along with precision, recall, and F1-scores of 82% (84% DR0, 82% DR1), 81% (85% DR0, 81% DR1), and 82% (84% DR0, 82% DR1), demonstrating its efficiency due to its lightweight structure. The results indicate that transfer learning on a combined dataset can effectively classify DR, though there remains room for improvement in distinguishing DR severity among 2 classes. In [16] the authors used the Messidor dataset for the classification of DR into two categories: normal and severe NPDR (Non-Proliferative Diabetic Retinopathy). Various CNN models, including VGGNet, AlexNet, InceptionNet, GoogleNet, DenseNet, and ResNet were evaluated. Image preprocessing techniques, such as image cropping and resizing, were applied to enhance image quality. The dataset was split into 70% for training, and 30% for testing. Among the tested models, ResNet-50 combined with SVM techniques achieved the highest overall testing accuracy of 95.83% for Base12, while InceptionV3 and VGGNet Type 19 achieved an overall testing accuracy of 95.24% for Base13. The study highlights that hybrid approaches, integrating CNN-based feature extraction with SVM, can improve DR classification performance. In [1] various CNN models, including AlexNet, VGG-s, VGGNet-vd-16, and VGGNet-vd-19, were utilized to classify the five DR severity levels. Image preprocessing techniques, such as image cropping and resizing, were applied to improve the quality of the images. Among the models, VGG-m demonstrated notable performance, achieving a specificity of overall 74.31%, overall sensitivity of 54.41%, overall accuracy of 65.6%, and an AUC of overall 0.7058. The study suggests that further improvements in preprocessing techniques and model architectures could enhance classification accuracy and robustness.However, the authors of this paper has not mentioned the per-class results and have only reported specificity, sensitivity, accuracy and AUC for the overall performance.

## II. Method

Fig. 1 illustrates the overall workflow of the proposed cascaded DR classification system. The process begins with training and testing images, which undergo pre-processing steps such as cropping, resizing to 224×224, and median filtering. A structured data augmentation strategy is then applied to enhance image diversity. The augmented images are fed into a pre-trained ResNet-50 model for initial classification. Predictions labelled as Severe or Proliferative DR are finalized at this stage, while those classified as No DR, Mild DR, or Moderate DR are further refined using the ResNet-50 model.

### A. Dataset

This study used two publicly available datasets to develop and evaluate the proposed cascaded DR classification framework for identifying the five severity levels of DR. The datasets used are APTOS 2019 [17], which consists of 3,662 images, and the Diabetic Retinopathy Resized Dataset [5], which contains 35,126 images. Both datasets classify DR into five distinct stages, ranging from 0 to 4, where 0 represents No DR, 1 indicates Mild DR, 2 corresponds to Moderate DR, 3 signifies Severe DR, and 4 denotes Proliferative DR. The distribution of images across these classes is summarized in Table I. Together, they offer an extensive and clinically relevant collection of retinal fundus images with annotations of DR severity levels. In order to conduct a balanced assessment and provide good training, the two datasets were provided and then divided into 80% training and 20% testing. The selection of datasets in our study is guided by prior research [2], which evaluated two DR datasets and found that combining the APTOS 2019 and Diabetic Retinopathy Detection datasets led to better classification performance. Based on these findings, we used both datasets together.

### B. Preprocessing

To ensure consistency and enhance the visual quality of the retinal fundus images, three preprocessing steps were applied prior to model training. First, a cropping technique [18] was used to eliminate black borders and center the retinal disc within each frame. This step helped reduce background noise and ensured that the model focused on the relevant retinal structures. Next, all images were resized to a fixed resolution of 224 × 224 pixels to match the input dimensions required

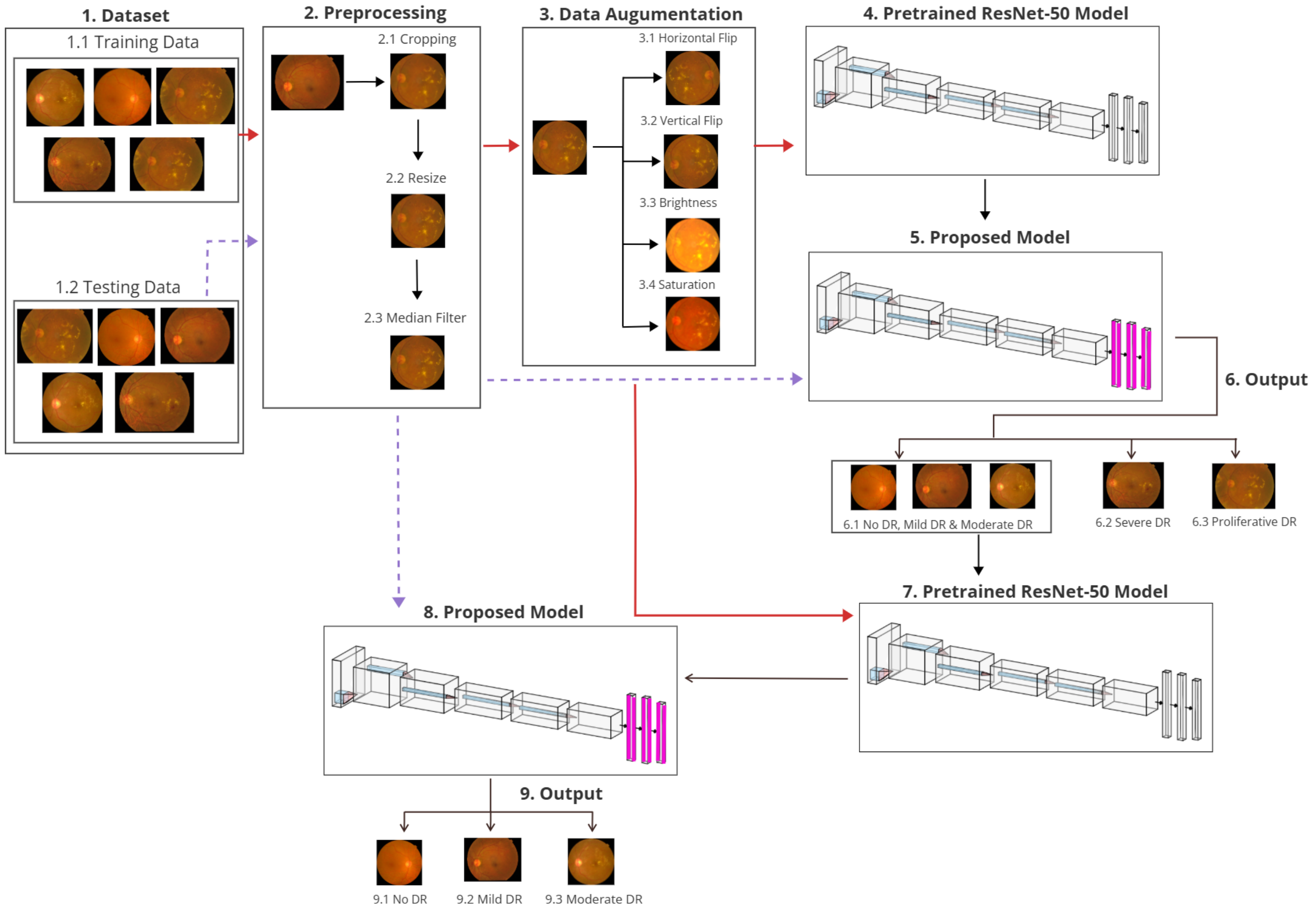

Fig. 1. Proposed cascaded approach for Diabetic Retinopathy (DR) classification: (1) dataset, (1.1) training data, (1.2) testing data, (2) pre-processing for training and testing dataset, (2.1) cropping out black portions, (2.2) resizing images to 224 × 224 pixels, (2.3) apply median filter to images, (3) data augmentation for training dataset, (3.1) horizontal flip to median filtered images, (3.2) vertical flip to median filtered images, (3.3) brightness enhancement, (3.4) saturation enhancement, (4) Training Images of 224 × 224 fed into pre-trained X model, (5) proposed model does classification from pre-processed testing data, (6) First model output classification, (6.1) No DR, Mild and Moderate DR, (6.2) Severe DR, (6.3) Proliferative DR, (7) training images from 6.2 fed into pre-trained Y model, (8) proposed model does classification of 6.2 testing images, (9) Second model output classification, (9.1) No DR, (9.2) Mild DR, (9.3) Moderate DR.

TABLE I
DISTRIBUTION OF DIABETIC RETINOPATHY SEVERITY LEVELS ACROSS DRC AND APTOS DATASETS

| Class | Severity Level | DRC | APTOS | Total |
|---|---|---|---|---|
| 0 | No DR | 1,805 | 25,810 | 27,615 |
| 1 | Mild DR | 370 | 2,443 | 2,813 |
| 2 | Moderate DR | 999 | 5,292 | 6,291 |
| 3 | Severe DR | 193 | 873 | 1,066 |
| 4 | Proliferative DR | 295 | 708 | 1,003 |
| **Total** | | **3,662** | **35,126** | **38,788** |

by the pre-trained CNN employed in this study. This standardization aided efficient batch processing and advantaged to the model convergence. Lastly, to reduce minor image noise and highlight small vascular structures, a 3 × 3 median filter [11] was used. This filtering process is particularly useful towards enhancing the clarity of microaneurysms and hemorrhages, which are very important diagnostic parameters in DR classification.

### C. Smooth Data Augmentation Strategy

To enhance model performance and address class imbalance, we adopted a data augmentation oversampling strategy, as supported by findings in [2], which showed improved results using this approach. Specifically, our study employs a deterministic and sequential method known as the Smooth Data Augmentation Strategy. The augmentation process begins with a horizontal flip, which mirrors the image along the vertical axis to introduce lateral variation [19]. Next, a vertical flip to simulate top-to-bottom structural differences [20]. In the third step, brightness is increased to reflect common variations in illumination during fundus imaging [21]. Finally, saturation is enhanced to account for contrast differences caused by varying imaging equipment or patient-specific factors [22]. These transformations are applied in a fixed sequence to ensure consistency while diversifying the training data. As a result, each original image is expanded into four distinct augmented versions. These augmented images are then passed into the first pre-trained model for initial classification. The final class-balanced image counts resulting from this augmentation process are summarized in Table II

TABLE II
DATASET DISTRIBUTION BEFORE AND AFTER SMOOTH DATA AUGMENTATION

| Stage | No DR | Mild | Moderate | Severe | PDR |
|---|---|---|---|---|---|
| Original Dataset | 27,615 | 2,813 | 6,291 | 1,066 | 1,003 |
| Training Before Aug. | 22,092 | 2,252 | 5,032 | 852 | 802 |
| Training After Aug. | 27,615 | 27,615 | 27,615 | 27,615 | 27,615 |
| Testing | 5,523 | 563 | 1,259 | 214 | 201 |

### D. Cascade-Based Ensemble Architecture

The cascade-based ensemble architecture is a hierarchical learning approach where multiple models are arranged in sequential stages, with each stage building upon the output of the previous one. This structure is particularly useful for complex classification problems, especially when certain inputs require more detailed analysis due to overlapping or subtle visual features. By processing data step by step, the cascade allows for more accurate and refined predictions, with each model in the sequence specializing in a specific part of the decision-making process.

In this study, four CNNs models were explored for their suitability in a cascaded framework, including **MobileNet** [23], **AlexNet** [24],**DenseNet** [25], and **ResNet50** [26], each known for their strengths in lightweight computation and effective feature extraction. After comparative evaluation, **ResNet50** was chosen in both cascade steps as the primary model due to its residual learning mechanism, which enables deeper network training by addressing issues like vanishing gradients. Additionally, using pre-trained versions of ResNet50 originally trained on large datasets such as ImageNet [27] helps transfer useful visual features to our specific task with minimal adjustment. The architectural details of the CNN models used in this study are provided in Table III.

## III. EVALUATION AND EXPERIMENT RESULTS

In this study, we developed a multi-stage cascaded classification framework to categorize DR into five severity levels. Initially, preprocessed and augmented images are passed through a pre-trained ResNet-50 model to predict all five DR classes. The outputs from this stage are then passed to a proposed CNN model, which focuses specifically on the early-stage predictions (classes 0, 1, and 2) while retaining the Severe and Proliferative DR classes (3 and 4). The advanced stages are finalized at this point, while the grouped early-stage predictions are passed through another ResNet-50 model for more precise classification. This multi-stage approach helps improve classification, particularly for early-stage DR, where differences are subtle.

### A. Evaluation Metrics

To evaluate the performance of the proposed models, this study uses standard classification metrics: Precision [28], Recall [28], and F1-score [29]. These metrics are particularly relevant in the context of DR diagnosis due to the class imbalance often observed in medical imaging datasets.

**Precision** measures the proportion of correctly predicted positive cases out of all cases predicted as positive. It reflects how precise or accurate the model's positive predictions are. It is defined as:

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

**Recall**, also known as Sensitivity or True Positive Rate, calculates the proportion of actual positive cases that were correctly identified by the model. It is given by:

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

**F1-score**is the harmonic mean of Precision and Recall, providing a balance between the two. It is especially useful when there is an uneven class distribution:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where TP is True Positive, correctly predicted positive cases; FP is False Positive, incorrectly predicted as positive when they are negative; and FN is False Negative, incorrectly predicted as negative when they are positive.

### B. Initial Classification of All Five DR Stages

In the initial round of the experiment, we trained all models to classify the five classes directly among the four architectures tested: ResNet50, DenseNet, MobileNet, and AlexNet. ResNet50 was superior in most of the metrics clear from Table IV and Fig. 2. The ResNet50 model achieved the highest overall F1-score of 0.64, along with strong performance in detecting the No DR class, with an F1-score of 0.79. While MobileNet recorded the highest F1-score if 0.84 for No DR, it showed limited effectiveness in other classes, particularly Mild DR, where the F1-score dropped to 0.08. AlexNet had a low overall F1-score of 0.60, and despite achieving the highest precision for Mild DR in earlier results, it failed to recall any cases, resulting in a very low F1-score of 0.01 for Mild DR. DenseNet, although slightly better on Mild DR with an F1-score of 0.15, did not surpass ResNet50 in any class. Overall, ResNet50 stood out for its balanced performance across multiple DR stages, including the highest F1-scores for Moderate DR (0.33), Severe DR (0.29), and Proliferative DR (0.43), making it the most consistent and reliable model for both early and advanced DR detection. Overall, ResNet50 provided the most consistent results across all classes, offering a better trade-off between precision and recall, especially in classifying both No DR and more severe DR conditions.

TABLE III
DETAILS OF THE ARCHITECTURES USED IN THIS STUDY

| Item | ResNet50 | AlexNet | DenseNet201 | MobileNetV2 |
|---|---|---|---|---|
| **General** | | | | |
| Parameters (approx) | 24.6M | 46.8M | 19.3M | 2.9M |
| Channels (Input) | 3 | 3 | 3 | 3 |
| Input size | 224×224 | 224×224 | 224×224 | 224×224 |
| **Number of layers** | | | | |
| Convolutional | 49 | 5 | 201 | 53 |
| Fully connected | 2 | 3 | 2 | 2 |
| **Presence of module** | | | | |
| Batch normalization | yes | yes | yes | yes |
| Residual connections | yes | no | no | yes |

TABLE IV
PRECISION AND RECALL FOR EACH DR CLASS AND OVERALL IN FIVE-CLASS CLASSIFICATION

| Class | ResNet50 | | DenseNet | | MobileNet | | AlexNet | |
|---|---|---|---|---|---|---|---|---|
| | Prec. | Rec. | Prec. | Rec. | Prec. | Rec. | Prec. | Rec. |
| No DR | **0.79** | 0.79 | 0.76 | 0.74 | 0.74 | 0.97 | 0.72 | **0.99** |
| Mild DR | 0.11 | 0.11 | 0.13 | 0.17 | **0.25** | 0.05 | **0.40** | 0.00 |
| Moderate DR | **0.33** | 0.33 | 0.22 | 0.17 | **0.41** | 0.11 | 0.33 | 0.02 |
| Severe DR | 0.29 | **0.29** | 0.09 | 0.20 | **0.30** | 0.03 | 0.29 | 0.01 |
| PDR | **0.50** | **0.38** | 0.15 | 0.22 | 0.18 | 0.05 | 0.22 | 0.07 |
| **Overall** | **0.64** | 0.64 | 0.59 | 0.57 | 0.62 | 0.71 | 0.60 | **0.71** |

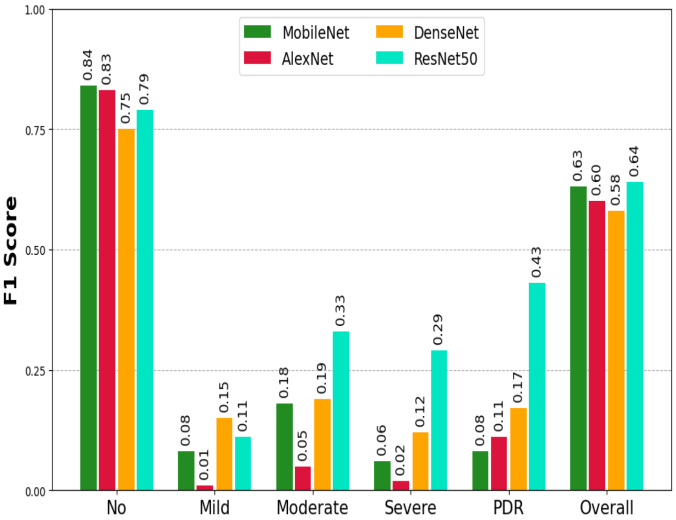

Fig. 2. F1-Scores for Five-Class Diabetic Retinopathy Classification.

### C. *Grouped Classification for Improved Classification.*

To improve classification performance and reduce misclassification between visually similar stages, we combined No DR (Class 0), Mild DR (Class 1), and Moderate DR (Class 2) into a single group because these early stages often present subtle differences that are difficult for the model to distinguish. However, Severe DR (Class 3) and Proliferative DR (Class 4) were kept as separate classes due to their more distinct and recognizable retinal features. This round aimed to simplify the decision boundaries in the first stage of the cascade, based on an exhaustive analysis of class-wise misclassifications observed in the confusion matrix. The results, shown in Table V and Fig 3, demonstrate considerably improved model performance. All models showed improved overall F1-scores in this stage due to better separation between early and advanced DR stages. For the grouped class (0,1,2), ResNet50 achieved a precision of 0.97, recall of 0.97, and an F1-score of 0.97, making it the most accurate and balanced model for identifying early-stage DR cases. MobileNet and AlexNet also performed well for the grouped class, each achieving an F1-score of 0.97, supported by precision and recall values of 0.91 and 0.94, respectively. However, in advanced stages, ResNet50 showed clear superiority. For Severe DR, it recorded an F1-score of 0.29, significantly higher than the other models, whose scores remained below 0.12. For Proliferative DR, ResNet50 again led with an F1-score of 0.43, backed by 0.50 precision and 0.38 recall. Overall, ResNet50 achieved the highest F1-score of 0.93, outperforming MobileNet and AlexNet both at 0.92 and DenseNet 0.89, demonstrating its strength in both grouped and advanced DR classification. In contrast, MobileNet and AlexNet, although comparable in overall F1-score, showed reduced reliability in precision and recall, particularly for severe and proliferative stages. This findings also validate that coarse grouping before fine classification helps the model generalize better to overlapping DR classes.

TABLE V
PRECISION AND RECALL FOR GROUPED EARLY-STAGE AND INDIVIDUAL ADVANCED DR CLASSES

| Class | MobileNet | | AlexNet | | DenseNet | | ResNet50 | |
|---|---|---|---|---|---|---|---|---|
| | Prec. | Rec. | Prec. | Rec. | Prec. | Rec. | Prec. | Rec. |
| Group (0,1,2) | 0.95 | **0.99** | 0.95 | 0.99 | 0.96 | 0.92 | **0.97** | 0.97 |
| Severe DR | **0.30** | 0.03 | 0.29 | 0.01 | 0.09 | 0.20 | 0.29 | **0.29** |
| PDR | 0.18 | 0.05 | 0.22 | 0.07 | 0.15 | 0.20 | **0.50** | **0.38** |
| **Overall** | 0.91 | **0.94** | 0.91 | 0.94 | 0.91 | 0.87 | **0.93** | 0.93 |

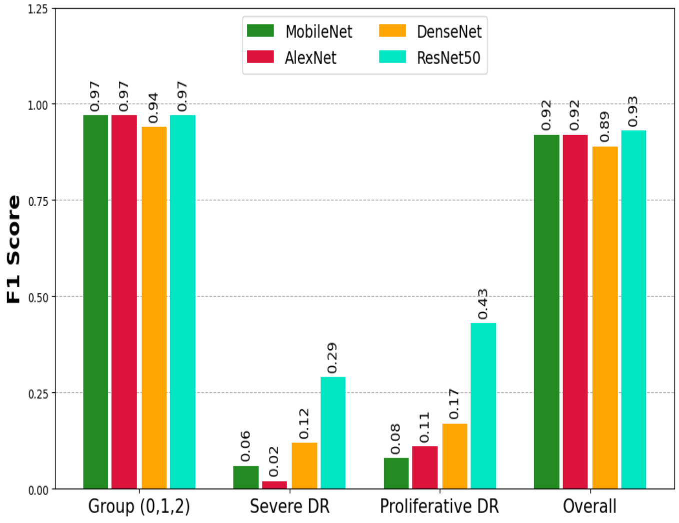

Fig. 3. F1-Score for Grouped and Advanced DR Classification.

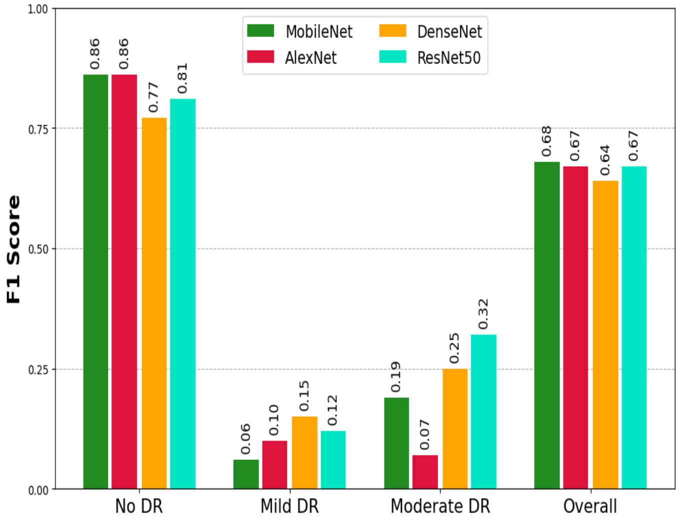

Fig. 4. F1-Score for No DR, Mild DR, and Moderate DR Classification

### *D. Refined Classification of Early DR Stages: No DR, Mild DR, and Moderate DR*

In the final round, only samples from the early DR group, No DR (Class 0), Mild DR (Class 1), and Moderate DR (Class 2) are used for further classification. The stages look similar to each other visually, but the cascade helped in narrowing the focus and improving detection within this group. As presented in Table VI and Fig. 4, For No DR, both MobileNet and AlexNet achieved the highest F1-scores 0.86, and MobileNet with 0.76 precision and 0.98 recall, on other hand AlexNet with 0.75 precision and 1.00 recall. However, performance dropped notably for Mild DR. DenseNet achieved the highest F1-score of 0.15, with a relatively better balance of 0.16 precision and 0.13 recall, while ResNet50 with F1-score of 0.12 and AlexNet with F1-score of 0.10, respectively. MobileNet performed the worst in this class. For Moderate DR Class 2, ResNet50 stood out by achieving the highest F1-score of 0.32, driven by 0.33 precision and 0.31 recall. Considering overall performance, ResNet50 demonstrated the most consistent balance across all three stages, making it better suited for real-world applications where early-stage DR detection is just as critical as identifying health cases.

TABLE VI
PRECISION AND RECALL FOR INDIVIDUAL EARLY DR STAGES: NO DR, MILD DR, AND MODERATE DR

| **Class** | **MobileNet** | | **AlexNet** | | **DenseNet** | | **ResNet50** | |
|---|---|---|---|---|---|---|---|---|
| | **Prec.** | **Rec.** | **Prec.** | **Rec.** | **Prec.** | **Rec.** | **Prec.** | **Rec.** |
| Class 0 | 0.76 | **0.98** | 0.75 | 1.00 | 0.78 | 0.77 | **0.78** | 0.84 |
| Class 1 | 0.26 | 0.04 | **0.28** | 0.06 | 0.13 | **0.16** | 0.13 | 0.10 |
| Class 2 | 0.42 | 0.13 | **0.61** | 0.04 | 0.24 | 0.27 | 0.33 | **0.31** |
| **Overall** | 0.66 | 0.76 | **0.69** | **0.76** | 0.64 | 0.64 | 0.65 | 0.69 |

### *E. Comparison with Literature*

This section compares our proposed model with selected state-of-the-art approaches that used similar datasets and deep learning techniques for DR classification. The goal is to evaluate performance differences, particularly in early-stage detection, where class overlap often presents significant challenges.

A study by [3] applied transfer learning using AlexNet and DenseNet-169 on the APTOS and Diabetic Retinopathy datasets. While DenseNet-169 reported F1-scores of 0.55 (No DR), 0.38 (Mild DR), and 0.40 (Moderate DR), our ResNet50 model achieved higher scores in most categories, including 0.81 for No DR and 0.32 for Moderate DR, demonstrating a more effective classification of clearly defined stages. Although the F1-score for Mild DR (0.12) was lower in our case, this class is widely recognized as the most challenging due to subtle visual cues and high inter-class overlap. Unlike the single stage classification approach used in the prior study, our approach incorporates a multi-stage cascaded framework, which allows for more refined decision-making across classes. Combined with a more diverse dataset, this design supports improved generalization and better performance on real-world and edge cases. The comparative results are visualized in Fig. 5, clearly reflecting the strength of our method across key DR stages. Although dataset size and testing setups may differ, our results provide clear evidence of improved class-wise performance and architectural robustness.

The study by [1] developed a ResNet50-based model for DR classification using only the Diabetic Retinopathy (resized) dataset and reported a high validation accuracy of 95.01%. However, as their evaluation was based on a single dataset with a fixed test ratio, a direct comparison with our model is not appropriate. Our study combines two publicly available datasets, introducing greater variability in image quality, class imbalance, and real-world complexity. Despite the increased

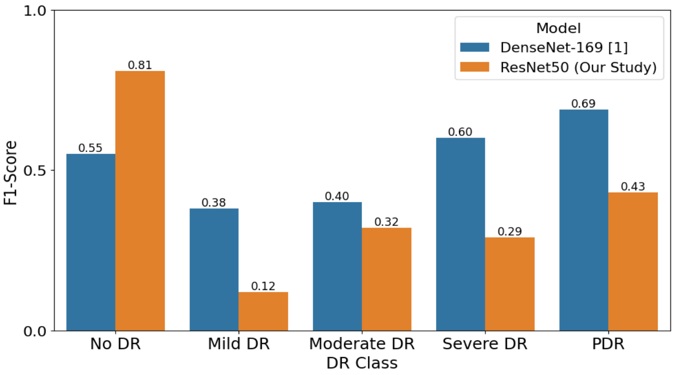

Fig. 5. F1-Score for Five-Class DR Classification.

difficulty of this setting, our model demonstrates stronger performance in detecting early-stage DR particularly Moderate DR, where it achieved the highest F1-score (0.32) among all models tested. This highlights the effectiveness of our cascaded design in capturing subtle pathological signs, which are often missed by traditional single-pass approaches. As shown in Figure 6, our model emphasizes robustness, generalization, and early-stage sensitivity key priorities for real-world screening systems.

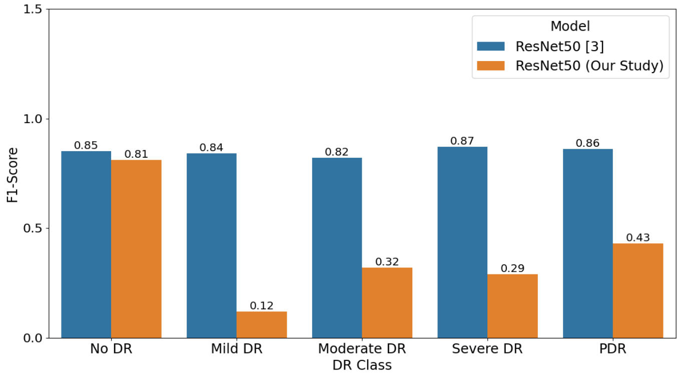

Fig. 6. F1-Score for Five-Class DR Classification.

According to [11] used ResNet50 with transfer learning on two datasets and reported strong overall results, precision 81.19% and recall 56.28%, but did not include any class-wise breakdown. Their evaluation was limited to overall performance, making it difficult to assess how well the model performs on specific DR stages, especially the early ones. In contrast, our study provides comprehensive overall metrics for ResNet50: Precision: 93%, Recall: 93%, and also provides detailed class-wise F1-scores. Furthermore, our study performs detailed classification across individual DR stages from Class 0 (No DR) to Class 2 (Moderate DR), which enhances the effectiveness of the model. This level of granularity highlights strengths and challenges in early-stage detection, offering a more complete and transparent evaluation of model performance.

In summary, unlike previous works that primarily report overall performance metrics, our study provides a more comprehensive evaluation by including both overall and class-wise results. This detailed analysis, especially for early DR stages, demonstrates our model's strength in classification. The use of a cascaded ResNet50 architecture along with two diverse datasets not only enhances classification accuracy but also ensures robustness and wider applicability, making our approach highly suitable for real-world DR screening applications.

## IV. Conclusions

In this study, we developed a cascaded DR classification framework to categorize DR into five severity levels: No DR, Mild DR, Moderate DR, Severe DR, and Proliferative DR, using two publicly available datasets: APTOS 2019 and the Diabetic Retinopathy Resized dataset. The system was designed to address challenges such as class imbalance and the visual similarity between early-stage DR classes. By grouping early and advanced stages during an intermediate classification step and refining predictions for closely related classes, our approach achieved notable improvements. Among the evaluated models, ResNet50 consistently performed the best, with an overall F1-score of 0.93 when Classes 0 to 2 (No DR, Mild DR, Moderate DR) were grouped, and 0.67 during the final classification stage of these three classes. Within this setting, No DR (Class 0) was the best performing class with an overall F1-score of 0.81, while Mild DR (Class 1) posed the most difficulty, achieving only 0.12 due to subtle visual differences. These findings confirm the effectiveness of our cascaded framework in improving accuracy, particularly in the early and visually ambiguous stages of DR.

Compared to previous research, the proposed cascaded ResNet50 framework offers clear improvements in class-wise diabetic retinopathy classification, particularly in the early stages. For instance, DenseNet-169 in [2] achieved an F1-score of 0.69 for Proliferative DR, but performed poorly on Mild and Moderate DR, with F1-scores of 0.38 and 0.40, respectively. Similarly, the ResNet50 model used in [6] showed limited sensitivity (56.52%) and specificity (85.71%) when not supported by a refined architecture. Additionally, earlier studies such as [4] and [23] often relied on binary classification or did not provide class-wise results, limiting their diagnostic value. In contrast, our model demonstrates more consistent performance across all five DR classes. It achieves significantly better F1-scores in detecting challenging stages like Moderate DR (0.33), Severe DR (0.29), and Proliferative DR (0.43), while also maintaining reliable results in earlier stages. This highlights the benefit of our multi-stage design in improving detection accuracy where early intervention is most critical, and further supports its potential for real-world DR screening applications.

Looking ahead, future work will focus on improving the interpretability of our cascade-based model using explainable AI techniques approaches tailored for classification tasks. We also aim to expand the dataset with more diverse and real-world DR images to enhance model robustness. Additionally, refining the cascaded architecture to better handle challenging cases particularly early-stage DR will be a key focus to support more reliable decision-making in screening environments. We also aim to optimize the model for lightweight deployment, enabling it to run efficiently on mobile or low-resource devices. This would facilitate early DR screening in remote areas where advanced equipment is not readily available.

## Disclaimer

Some sentences in this manuscript were simplified and refined using Grammarly and GenAI tools to improve grammar, clarity, and semantic correctness. GenAI tools were also used for code debugging support.

## References

[1] G. Sharma, V. Anand, R. Chauhan, H. S. Pokhariya, S. Gupta, and G. Sunil, "Multiclass identification and classification of diabetic retinopathy using transfer learning: A comprehensive approach for enhanced diagnostic," in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, 2024, pp. 1–6.

[2] S. H. Kassani, P. H. Kassani, R. Khazaeinezhad, M. J. Wesolowski, K. A. Schneider, and R. Deters, "Diabetic retinopathy classification using a modified xception architecture," in *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2019, pp. 1–6.

[3] R. Baskar, E. Sabu, and C. Mazo, "Deep cnns for diabetic retinopathy classification: A transfer learning perspective," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024, pp. 1–4.

[4] X. Li, T. Pang, B. Xiong, W. Liu, P. Liang, and T. Wang, "Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification," in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2017, pp. 1–11.

[5] A. P. T.-O. S. (APTOS), "Aptos 2019 blindness detection dataset," 2019, accessed: 2024-12-30. [Online]. Available: https://www.kaggle.com/competitions/aptos2019-blindness-detection/data

[6] S. Dutta, B. Manideep, S. M. Basha, R. Caytiles, and N. C. S. N. Iyenger, "Classification of diabetic retinopathy images by using deep learning models," *International Journal of Grid and Distributed Computing*, vol. 11, pp. 89–106, 01 2018.

[7] N. Khalifa, M. Loey, M. Taha, and H. Mohamed, "Deep transfer learning models for medical diabetic retinopathy detection," *Acta Informatica Medica*, vol. 27, no. 5, pp. 327–332, Dec 2019.

[8] M. Al-Smadi, M. Hammad, Q. B. Baker, and A. Sa'ad, "A transfer learning with deep neural network approach for diabetic retinopathy classification," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 4, p. 3492, 2021.

[9] P. K. Das and S. Pumrin, "Cnn transfer learning for two stage classification of diabetic retinopathy using fundus images," in *2023 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT NCON)*, 2023, pp. 443–447.

[10] D. U. N. Qomariah, H. Tjandrasa, and C. Fatichah, "Classification of diabetic retinopathy and normal retinal images using cnn and svm," in *2019 12th International Conference on Information Communication Technology and System (ICTS)*, 2019, pp. 152–157.

[11] H. Gupta, N. Arora, S. Dutt, A. Gulati, and A. Gulati, "Detection of diabetic retinopathy using transfer learning," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2023, pp. 1–8.

[12] S. Tyagi, S. Pingulkar, and A. Tiwary, "Detecting diabetic retinopathy using resnet50 and explainable ai," in *2023 IEEE International Carnahan Conference on Security Technology (ICCST)*, 2023, pp. 1–6.

[13] I. Aiche, Y. Brik, B. Attallah, H. Lahmar, and Z. Zohra, "Transfer learning for diabetic retinopathy detection," in *2022 International Conference of Advanced Technology in Electronic and Electrical Engineering (ICATEEE)*, 2022, pp. 1–5.

[14] P. Verma and S. Elango, "Improving diabetic retinopathy diagnosis with transfer learning and filtered fundus images: An automated approach," in *2023 International Conference on Modeling, Simulation Intelligent Computing (MoSICom)*, 2023, pp. 456–461.

[15] S. Dasari, B. Poonguzhali, and M. Rayudu, "Transfer learning approach for classification of diabetic retinopathy using fine-tuned resnet50 deep learning model," 11 2023, pp. 1361–1367.

[16] ILOVESCIENCE, "Diabetic retinopathy - resized," 2018, accessed: 2024-12-30. [Online]. Available: https://www.kaggle.com/datasets/tanlikesmath/diabetic-retinopathy-resized

[17] V. Vipparthi, D. R. Rao, S. Mullu, and V. Patlolla, "Diabetic retinopathy classification using deep learning techniques," in *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2022, pp. 840–846.

[18] Wikipedia contributors, "Cropping (image)," https://en.wikipedia.org/wiki/Cropping%28image%29, 2025, accessed: 2025-07-17.

[19] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[20] X. Wang and L. Perez, "The effectiveness of data augmentation in image classification using deep learning," *Convolutional Neural Networks Vis. Recognit*, vol. 11, no. 2017, pp. 1–8, 2017.

[21] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," in *NeurIPS Workshop on Machine Learning for Health*, 2017.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," in *arXiv preprint arXiv:1704.04861*, 2017.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems (NeurIPS)*, vol. 25, 2012, pp. 1097–1105.

[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017, pp. 4700–4708.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[28] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[29] D. M. W. Powers, "Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.