Automated Diabetic Retinopathy Grading Using Transfer Learning CNNs with Grad-CAM

Yasir Alalawi & Ajab Aldosari

447540166 & 447540089

Master in Artificial Intelligence

Computer Science Department

Prince Sattam bin Abdulaziz University

April 2026

Abstract

Diabetic retinopathy is a serious complication of diabetes, and early grading from retinal fundus images can help support screening and follow-up decisions. In this project, we built a deep learning pipeline to classify retinal images into five severity levels: No DR, Mild, Moderate, Severe, and Proliferative DR. Because the dataset is limited and imbalanced, we used transfer learning instead of training a convolutional neural network from scratch. ResNet50 was used as a reference model, and EfficientNetB0 was used as an efficient comparison model. In the final notebook run, both models were fine-tuned and ResNet50 achieved the stronger validation result, with 0.6903 accuracy, 0.5250 macro F1-score, and 0.8004 quadratic weighted kappa (QWK). EfficientNetB0 reached 0.6576 accuracy, 0.4889 macro F1-score, and 0.7931 QWK. We also applied Grad-CAM to visualize the image regions that influenced selected predictions. The results show that transfer learning is useful for this medical image task, but minority classes, especially Severe and Proliferative DR, remain difficult because of class imbalance and visual similarity between disease stages.

1 Introduction

Diabetic retinopathy (DR) is a retinal disease caused by long-term diabetes-related damage to the small blood vessels in the retina. If it is not detected and monitored early, it can lead to serious vision problems and, in advanced cases, vision loss. In real clinical practice, retinal fundus images are examined by specialists, but manual grading can take time and may not always be available in every setting. This was one reason we chose this topic, because it connects a real medical image problem with the deep learning methods studied in the course.

Our motivation for this problem became clear while we were reviewing the dataset and the grading task. A fundus image can show visible signs such as microaneurysms, hemorrhages, exudates, and other retinal changes. However, the difference between neighboring disease stages is not always obvious, especially for a non-specialist. A deep learning system is not meant to replace a physician, but it can support screening by quickly identifying images that may need more attention. This makes the problem suitable for a semester project because it has both practical medical meaning and a clear image classification structure.

In this project, our goal was to classify retinal fundus images into five ordered DR classes: No DR, Mild, Moderate, Severe, and Proliferative DR. The problem is challenging for two

main reasons. First, the dataset is imbalanced, with many more images in the No DR class than in the Severe class. Second, the classes are ordinal and visually related. A Severe case is not completely separate from a Moderate case; rather, the disease progresses through stages. Because of that, errors between neighboring grades are more understandable than random errors across distant classes.

For this project, we used pretrained convolutional neural networks (CNNs). The main idea was to use knowledge learned from large image datasets and adapt it to retinal image grading. This approach was more practical for our project because training a deep CNN from the beginning would require more data, more time, and more computing power. Transfer learning is also common in medical image analysis when the available dataset is smaller than the datasets used to train modern CNN backbones.

We compared two CNN architectures: ResNet50 as a reference model and EfficientNetB0 as an efficient transfer learning model. ResNet50 is a well-known residual network that is often used as a strong baseline in image classification tasks [1]. EfficientNetB0 is a lighter architecture that balances network depth, width, and image resolution more efficiently [2]. In addition to classification, we used Grad-CAM to inspect selected predictions visually [3]. This explainability step is important because, in a healthcare-related task, it is not enough to report only a number such as accuracy. We also need to ask whether the model appears to focus on meaningful image regions.

The rest of this paper is organized as follows. Section 2 reviews related work on deep learning for diabetic retinopathy detection and grading. Section 3 describes the dataset used in this project. Section 4 explains the methodology, including preprocessing, augmentation, class weighting, transfer learning, and evaluation metrics. Section 5 presents the results and model comparison. Section 6 discusses Grad-CAM explainability. Section 7 discusses the findings and future improvements. Section 8 summarizes the main limitations, and Section 9 concludes the paper.

2 Literature Review

2.1 Deep Learning for Diabetic Retinopathy Screening

Deep learning has become one of the main approaches for automated diabetic retinopathy analysis from fundus images. Earlier computer-aided methods often depended on manually designed features, such as detecting lesions or measuring vessel-related properties. Modern CNNs changed this process by learning image features directly from data. This is useful for DR because the important signs may appear in different shapes, sizes, and locations across images.

One of the important early studies in this area was the work by Gulshan et al. [4]. They developed and validated a deep learning algorithm for detecting diabetic retinopathy in retinal fundus photographs. That study was important because it showed that deep learning could reach strong performance on large retinal datasets and could be evaluated in a clinically meaningful way. Ting et al. [5] also developed a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations. Their work is relevant because it highlighted the need to test models on diverse data rather than relying only on one narrow dataset.

Other studies also showed the value of deep learning for DR screening. Gargeya and Leng [6] presented an automated identification approach using deep learning, while Abràmoff et al. [7] reported a pivotal trial of an autonomous AI diagnostic system. These studies are larger and more clinically oriented than our semester project, but they help explain why the topic is

important. They also show that strong validation and careful reporting are necessary before any model can be considered for real clinical use.

More recent review studies confirm that deep learning is now widely used for DR detection and grading. Tsiknakis et al. [8] reviewed deep learning methods for DR detection and classification based on fundus images. Their review discussed common datasets, preprocessing methods, classification pipelines, and lesion localization. Zhelev et al. [9] reviewed the test accuracy of AI-based grading systems for DR screening, which is useful because it reminds us that high accuracy in one experiment does not automatically mean a model is ready for clinical deployment. Naz et al. [10] reviewed supervised and unsupervised deep learning approaches for DR detection, while Ul Haq et al. [11] focused on computationally efficient deep learning models.

For our project, these studies provided two important lessons. First, CNN-based transfer learning is a reasonable approach for retinal image classification. Second, evaluation should not depend only on accuracy. Since DR datasets are often imbalanced, macro precision, macro recall, macro F1-score, confusion matrices, and class-wise reports are needed to understand how the model behaves across all disease grades.

2.2 Transfer Learning and Efficient CNN Architectures

Transfer learning is especially useful in medical image projects because collecting and labeling medical data is expensive. In many student projects, the available dataset is not large enough to train a deep CNN from scratch. A pretrained model gives a better starting point because it has already learned general visual features such as edges, textures, and shapes. These features are not specific to retinal images, but they can still help the model learn the new task faster.

ResNet introduced residual connections to make deep networks easier to train [1]. Because of this, ResNet50 is often used as a baseline model. In our project, ResNet50 was useful because it gave us a reference point for comparison. EfficientNet was proposed later with a compound scaling method that balances depth, width, and resolution [2]. EfficientNetB0 is the smallest version in the EfficientNet family, but it can still perform strongly while using fewer parameters than many older CNN architectures.

Several recent DR studies have used EfficientNet or related efficient architectures. Vijayan and Venkatakrishnan [13] proposed an EfficientNet-based approach for DR diagnosis. Batool et al. [14] studied EfficientNet-based grading of DR severity levels from fundus images. Alwakid et al. [16] also explored deep learning for DR classification, while Mane et al. [18] studied transfer learning deep neural networks for DR recognition and classification. These studies support the choice of testing an EfficientNet-based model in our project.

Other recent work also shows that researchers are still improving DR grading pipelines rather than treating the problem as fully solved. Alam et al. [15] studied contrastive pretraining for improving representation and transferability in DR classification models. Malhi et al. [17] discussed detection and grading using digital retinal images, and Shakibania et al. [20] proposed a dual-branch deep learning network for detection and stage grading. These works are more advanced than our implementation, but they are useful because they show the same general direction: better feature learning, better grading, and more careful evaluation.

More recent studies from 2024 to 2026 also show that APTOS 2019 is still being used as a benchmark for DR grading. Chilukoti et al. [19] used transfer learning and ensemble learning with the quadratic weighted kappa (QWK) metric on EyePACS, APTOS, and Messidor. Shakibania et al. [20] evaluated a dual-branch model on APTOS 2019 and reported strong stage-grading performance. Vallukappully et al. [21] studied NASNet-Large and ResNet50 with

class balancing and preprocessing tests on APTOS 2019. Dharrao et al. [22] proposed an EfficientNetB0-based model with multi-scale spatial attention gates, while Al-Dolat et al. [24] used a CheXNet-CBAM model with Grad-CAM on APTOS 2019 and DDR. Recent transformer-based studies also explored lightweight transformer models and attention-enhanced ViT models for DR grading [23, 26].

At the same time, transfer learning is not a magic solution. If the dataset is strongly imbalanced or if minority classes have few examples, a model can still learn biased decision patterns. This is why our methodology included class weights and macro metrics. The main purpose was not only to increase overall accuracy, but also to make the evaluation more fair across the five classes.

2.3 Class Imbalance in DR Grading

Class imbalance is one of the most common problems in DR datasets. In many real datasets, there are more normal or mild cases than severe disease cases. This makes sense from a population point of view, but it creates a machine learning problem. If a model sees many more No DR images than Severe images, it may learn to favor the majority class. A high accuracy score can then hide weak performance on the classes that matter clinically.

In our dataset, the Severe class has only 193 images out of 3,662. After the train-validation split, the validation set contains only 39 Severe images. This is a small number for evaluating a five-class classification model. Because of this, a few wrong predictions can strongly affect the Severe-class precision, recall, and F1-score. This is one reason we reported macro F1-score and class-wise results.

Recent reviews also discuss the imbalance problem in DR grading [8, 9, 10]. Some researchers use data augmentation, class weighting, resampling, focal loss, or ordinal classification methods to reduce the problem. In our project, we used class weights and light augmentation because they were simple, suitable for the project scope, and did not require changing the dataset itself. This was an honest and practical choice for a semester project.

2.4 Explainability in Medical Image Classification

Explainability is an important topic in medical image classification. A CNN can produce a class prediction, but without explanation it is difficult to know whether the prediction is based on meaningful retinal regions or on irrelevant image artifacts. Grad-CAM is a common method for visualizing which areas of an image contributed most to a CNN prediction [3]. It creates a heatmap that can be overlaid on the original image.

For diabetic retinopathy, explainability does not prove that a model is clinically correct. A heatmap can be useful, but it is not the same as a medical diagnosis. Still, it helps us inspect whether the model is focusing on retinal areas that seem relevant rather than background borders or unrelated regions. In our project, Grad-CAM was used as a visual check of model behavior. This fits the course goal because it connects CNN classification with model interpretation.

Recent work on deep learning for DR continues to include explainability or interpretation as an important direction [12, 11, 25]. For a student project, including Grad-CAM is useful because it shows that the project is not only about training a model and reporting accuracy. It also considers how the model output can be inspected and discussed critically.

2.5 Position of This Project Compared With Previous Work

Compared with large clinical studies, this project is smaller in scale. It uses a train-validation split rather than an independent external test set, and it reports transfer learning results instead of a fully optimized clinical-grade system. Even so, it still follows a solid deep learning workflow, including data verification, preprocessing, augmentation, class weighting, transfer learning, model comparison, class-wise evaluation, confusion matrix analysis, and Grad-CAM visualization.

For us, the value of this project was not only in the final accuracy, but also in understanding the full deep learning workflow. From an educational point of view, it shows how CNN transfer learning can be applied to a real medical image problem. From a practical point of view, it highlights the strengths and weaknesses of the approach on an imbalanced five-class dataset. We do not claim that the model is ready for medical use. Instead, we present it as a deep learning semester project that examines model behavior honestly and makes its limitations clear.

2.6 Lessons From the Literature for This Project

The literature also helped us understand what not to claim. Some published systems report very high performance, but many of them rely on larger datasets, external validation, ensemble models, or clinical grading protocols. Our project did not include those components, so it would not be fair to compare our validation accuracy directly with large clinical studies as if the experimental settings were the same. Instead, we used the literature as background support for the project design.

From the reviewed work, we took four main lessons. First, transfer learning is a reasonable strategy for fundus image classification when training data is limited. Second, class imbalance should be discussed clearly because it can hide weak performance on minority disease stages. Third, explainability should be treated as a way to inspect model behavior, not as proof that the model is clinically correct. Fourth, comparison with previous studies should include the experimental setting, not only the final number, because a model trained with ensemble learning, extra preprocessing, or an external test set is not directly equivalent to a course-scale transfer learning experiment. These points shaped the final report and helped us keep the writing honest.

3 Dataset

The dataset contains 3,662 retinal fundus images and a CSV file with image IDs and diagnosis labels. The labels follow five ordinal DR classes, as shown in Table 1. The images were stored in class folders, and all image paths were checked before training. This verification step was important because missing files or incorrect paths can cause silent mistakes later in the pipeline.

Table 1: Diabetic retinopathy classes used in the project.

Class ID	Class Name
0	No DR
1	Mild
2	Moderate
3	Severe
4	Proliferative DR

The class distribution is not balanced. Almost half of the images belong to the No DR class,

while the Severe class has the smallest number of images. This imbalance is important because a model can perform well on the majority class but still perform weakly on the minority disease stages.

Table 2: Class distribution in the dataset.

Class	Count	Percentage
No DR	1805	49.29%
Mild	370	10.10%
Moderate	999	27.28%
Severe	193	5.27%
Proliferative DR	295	8.06%

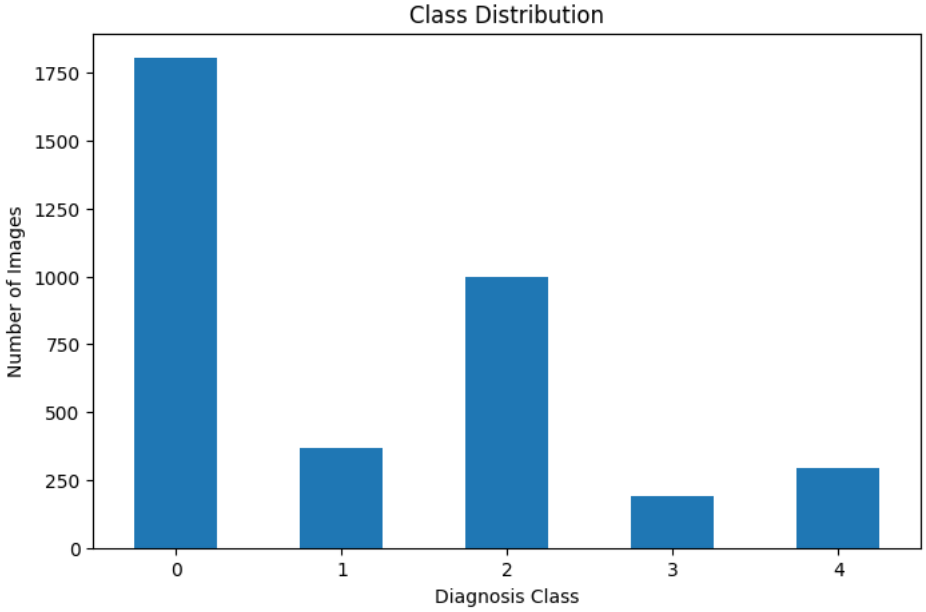


Figure 1: Class distribution of the diabetic retinopathy dataset.

Figure 1 makes the imbalance easier to see. The No DR class dominates the dataset, followed by Moderate. Severe and Proliferative DR have much smaller counts. This matters because the model receives fewer examples of the advanced disease stages during training.

Some sample images from the five classes are shown in Figure 2. Looking at the images helps explain why the task is difficult: the disease changes can be subtle, and not every class has an obvious visual difference for a non-expert. In addition, fundus images can differ in illumination, contrast, crop, and image quality.

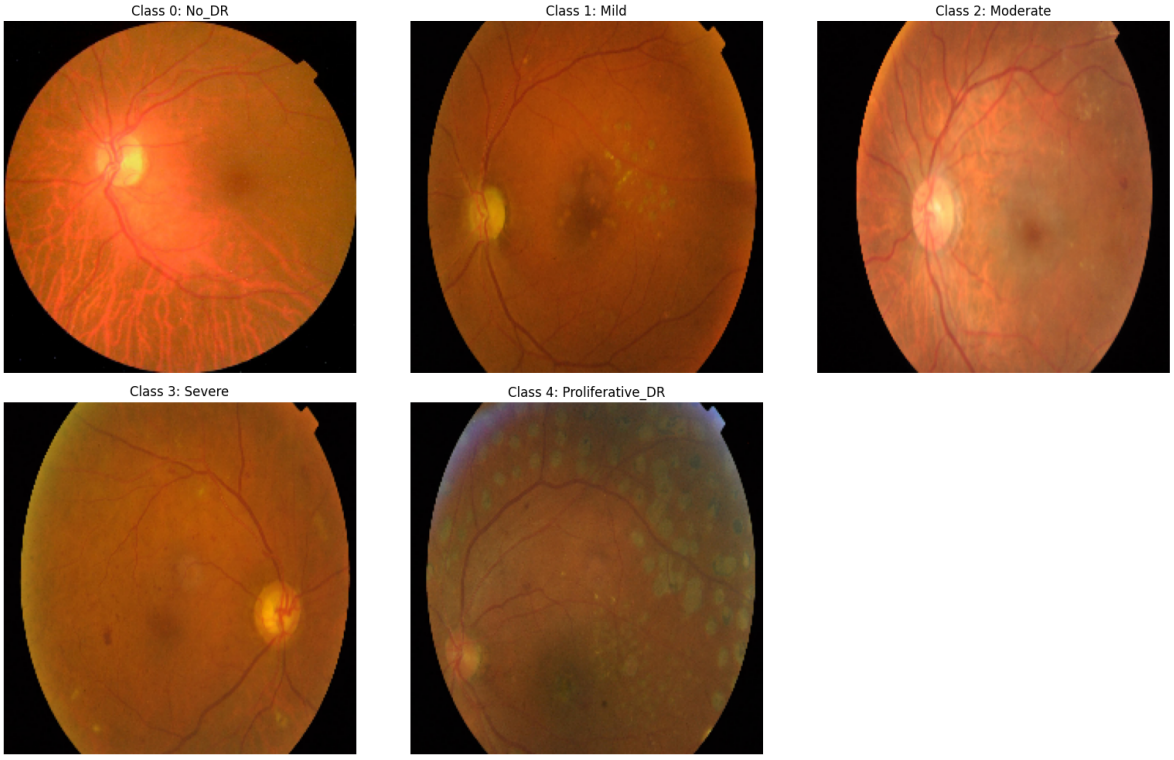


Figure 2: Sample retinal images from the five DR classes.

4 Methodology

4.1 Project Workflow

The overall workflow of the project was designed to be simple, reproducible, and suitable for the available dataset. First, the metadata file was loaded and matched with the image folder. Second, the images were split into training and validation sets using a stratified split. Third, image generators were prepared with preprocessing and augmentation. Fourth, two transfer learning models were trained and evaluated. Finally, Grad-CAM was used to visualize selected model predictions.

This workflow was selected because it follows a common deep learning pipeline without making the project too complicated. The purpose was not to build the most advanced possible DR system, but to implement a correct and explainable deep learning experiment that can be defended in a course project.

4.2 Train and Validation Split

The dataset was split into 80% training and 20% validation using a stratified split. Stratification keeps the class proportions similar in both sets. This is important because a random split without stratification could accidentally place too few minority-class examples in the validation set. The training set contains 2,929 images and the validation set contains 733 images.

Table 3: Train and validation split.

Class	Train Count	Validation Count
No DR	1444	361
Mild	296	74
Moderate	799	200
Severe	154	39
Proliferative DR	236	59

The split shows the main challenge clearly. The validation set has 361 No DR images but only 39 Severe images. This means that accuracy can be misleading. A model can obtain a reasonable accuracy by performing well on No DR and Moderate while still struggling with Severe and Proliferative DR. For this reason, the evaluation section includes macro metrics and confusion matrices.

4.3 Preprocessing

All images were resized to 224×224 pixels before being passed to the models. This size was used because it is compatible with the pretrained CNN backbones used in the project. Pixel values were also rescaled before training. These preprocessing steps are standard in transfer learning pipelines because CNNs expect input images to have consistent dimensions and scale.

The validation data was not augmented. It was only rescaled. This is important because validation should represent the model's performance on fixed images, not on randomly changed versions of those images. If validation images are randomly augmented, the evaluation can become unstable and harder to interpret.

4.4 Data Augmentation

The training generator used light augmentation, including rotation, small width and height shifts, zoom, and horizontal flipping. The purpose of augmentation was to expose the model to small variations of the training images. This can help reduce overfitting, especially when the dataset is not very large.

We kept the augmentation moderate. Strong augmentation could create images that are unrealistic for retinal diagnosis or change important details. Since the dataset is medical, it is better to use conservative transformations that do not distort the disease patterns too aggressively. This choice also made the experiment easier to explain and defend.

4.5 Handling Class Imbalance

Class weights were computed from the training labels and used during training. The minority classes received larger weights, especially the Severe class. This tells the loss function to give more importance to mistakes on underrepresented classes.

Using class weights does not completely solve imbalance, but it is a reasonable first step. It avoids simply duplicating images, and it keeps the original data split unchanged. In our project, class weighting was especially important because Severe and Proliferative DR had much fewer training examples than No DR and Moderate.

4.6 Transfer Learning Models

We tested two pretrained CNN models. ResNet50 was used as a reference model because it is a well-known residual CNN architecture. EfficientNetB0 was used as an efficient comparison model because it is lighter and more parameter-efficient than ResNet50.

For both models, the ImageNet pretrained backbone was loaded without the original classification layer. A new classification head was added for the five DR classes. During the first training stage, the pretrained backbone was frozen and only the custom head was trained. After that, both models were fine-tuned by unfreezing the upper layers and continuing training with a smaller learning rate. This made the final comparison cleaner because both models had the same general two-stage training structure: frozen-backbone transfer learning followed by limited fine-tuning.

4.7 Evaluation Metrics

The models were evaluated using accuracy, macro precision, macro recall, macro F1-score, quadratic weighted kappa (QWK), confusion matrices, and class-wise reports. Accuracy measures the total proportion of correct predictions, but it can be affected by class imbalance. Macro metrics calculate performance for each class and then average across classes equally. This gives a more honest view of performance on minority classes.

The macro F1-score was especially useful in this project. It balances precision and recall and gives equal importance to each class. Since the project is about five disease grades, we wanted to know whether the model works across all classes, not only on the majority class. QWK was also useful because DR grades are ordered. A prediction of Moderate instead of Severe is still wrong, but it is not as severe as predicting No DR for a Proliferative DR case. QWK reflects this ordered structure by giving larger penalties to errors that are farther from the true class.

4.8 Implementation Notes

The implementation was completed in a Jupyter notebook using TensorFlow/Keras. The notebook was organized so that each major step could be checked separately: importing libraries, verifying image paths, preparing the dataframe, creating data generators, training the models, evaluating the results, and generating Grad-CAM visualizations. This structure helped reduce mistakes because each stage produced output that could be inspected before moving to the next one.

One practical issue in this project was training time. ResNet50 required more time to train than EfficientNetB0 on the same data. The project therefore considered not only validation performance, but also practicality. In this final run, ResNet50 gave the stronger validation result, while EfficientNetB0 remained attractive from a training-time and efficiency point of view.

We also kept the final reporting connected to the executed notebook outputs. This is important because notebooks can contain exploratory cells that are not part of the final reported result. In the final notebook used for this report, both fine-tuning sections were executed, and the paper reports the saved evaluation outputs from that run. This avoids the common mistake of describing an experiment more strongly than the saved results support.

4.9 Why ResNet50 and EfficientNetB0 Were Compared

The comparison between ResNet50 and EfficientNetB0 was useful because the two models represent different design ideas. ResNet50 is deeper and uses residual connections, which help information flow through the network. EfficientNetB0 is designed to be more parameter-efficient by scaling network dimensions in a balanced way. For this dataset, EfficientNetB0 was expected to be a good candidate because it can learn useful image features without being too heavy.

Using two models also made the project stronger than reporting only one model. If only EfficientNetB0 had been trained, it would be harder to judge whether the result was reasonable. ResNet50 gave a baseline for comparison. In the final run, the two models produced close validation results. ResNet50 gave the stronger final validation result, while EfficientNetB0 remained useful as a lighter comparison model. This result was important because it showed that the more efficient architecture did not automatically perform better on this particular training run.

5 Results

5.1 ResNet50 Reference Model

The ResNet50 reference model reached a validation accuracy of 0.6903, a macro F1-score of 0.5250, and a QWK score of 0.8004. In the final notebook run, this result corresponds to the saved ResNet50 checkpoint after the short fine-tuning stage. The fine-tuning curves are shown in Figure 3. The model performed very well on the No DR class, but it still struggled with smaller classes such as Severe.

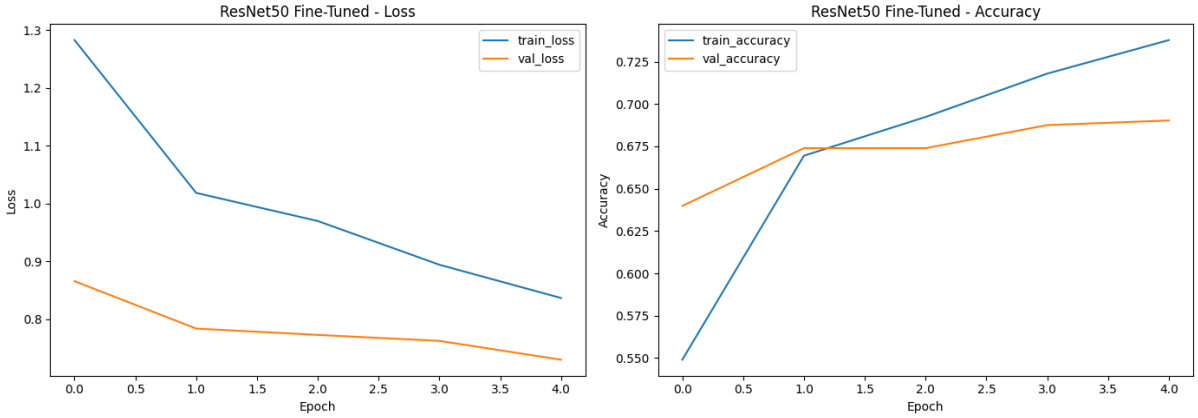


Figure 3: Fine-tuning training and validation curves for the ResNet50 reference model.

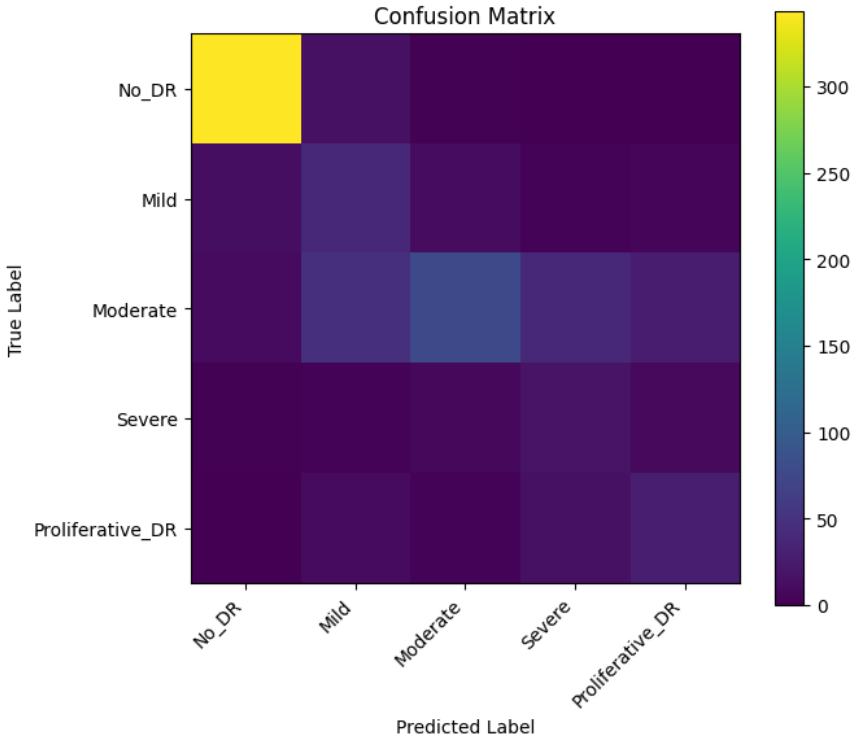


Figure 4: Confusion matrix for the ResNet50 reference model.

The ResNet50 confusion matrix in Figure 4 shows that No DR was classified correctly most of the time. However, many Moderate images were confused with Mild, Severe, or Proliferative DR. The class-wise results show that ResNet50 achieved a high F1-score for No DR, but the F1-score for Severe was only 0.31. This confirms that accuracy alone is not enough for this dataset.

5.2 EfficientNetB0 Comparison Model

In the final validation results, EfficientNetB0 achieved a validation accuracy of 0.6576, a macro F1-score of 0.4889, and a QWK score of 0.7931 after fine-tuning. The training curves are shown in Figure 5, and the confusion matrix is shown in Figure 6.

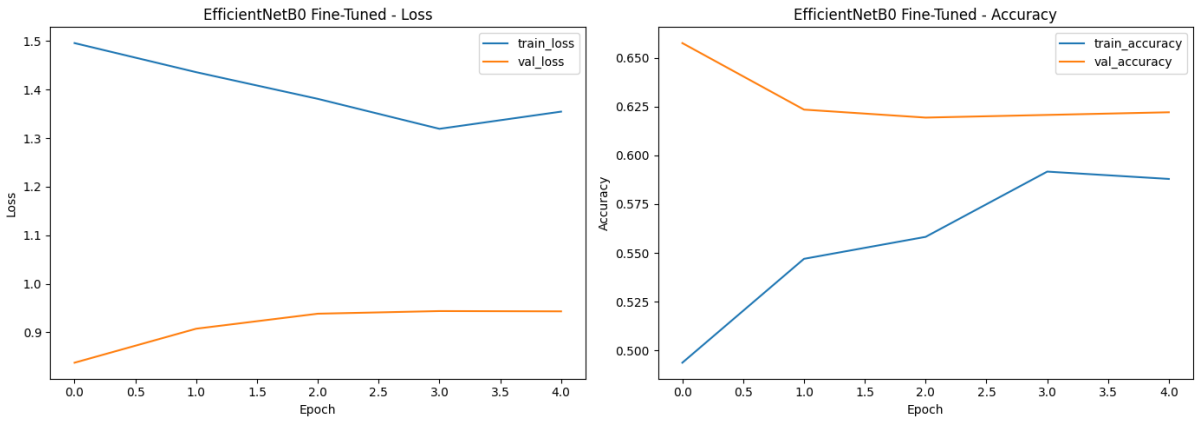


Figure 5: Fine-tuning training and validation curves for the EfficientNetB0 model.

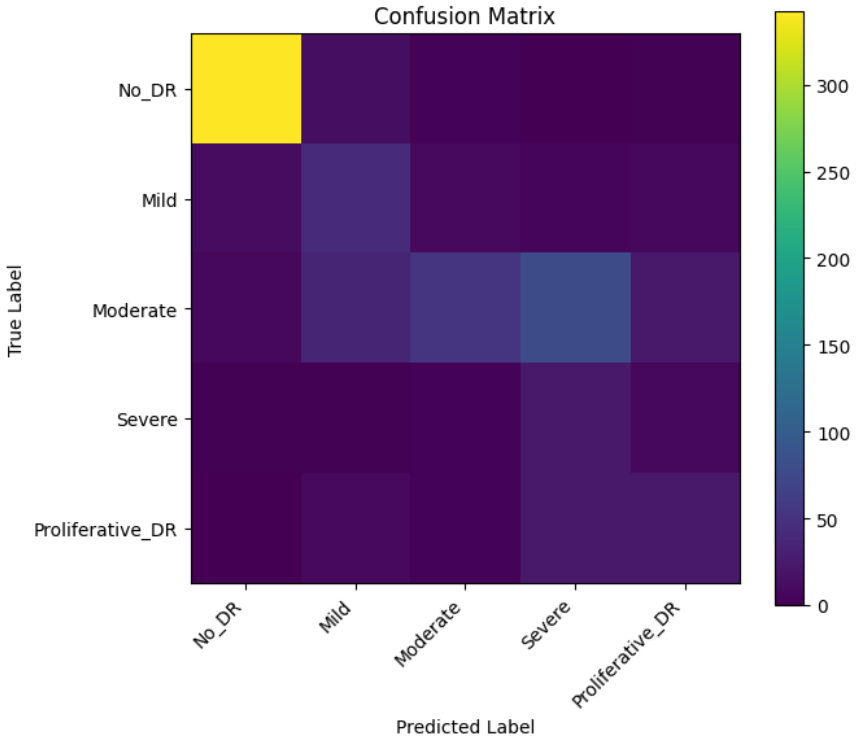


Figure 6: Confusion matrix for the EfficientNetB0 model.

The EfficientNetB0 model performed well on the No DR class, but its final overall metrics were lower than ResNet50. The Mild class F1-score was 0.47, the Moderate class F1-score was 0.39, the Severe class F1-score was 0.27, and the Proliferative DR F1-score was 0.37. These results show that the model learned useful patterns, but the smaller disease classes and visually similar grades remained difficult.

5.3 Model Comparison

Table 4 summarizes the main validation results. In the final run, ResNet50 achieved higher accuracy, macro precision, macro recall, macro F1-score, and QWK than EfficientNetB0. The difference was not extremely large, but it was consistent across the reported summary metrics.

Table 4: Validation performance comparison between ResNet50 and EfficientNetB0.

Model	Accuracy	Macro Precision	Macro Recall	Macro F1	QWK
ResNet50	0.6903	0.5359	0.5647	0.5250	0.8004
EfficientNetB0	0.6576	0.5215	0.5492	0.4889	0.7931

Based on these validation results, ResNet50 was the stronger model in the final notebook run. Its QWK score was also higher, which is useful because QWK considers the ordered nature of DR grades. Still, the macro F1-score is much lower than the accuracy for both models. This means the models are still biased toward easier or more common classes, especially No DR and Moderate. We noticed that accuracy looked acceptable, but macro F1 gave a more honest view of the minority-class problem.

5.4 Comparison with Recent Related Work

To address the broader academic context, we also compared our result with recent diabetic retinopathy grading studies, especially studies that used APTOS 2019 or closely related fundus-image datasets. This comparison is shown in Table 5. The purpose is not to claim that all numbers are directly equivalent. Different papers use different train-test splits, preprocessing methods, balancing strategies, model sizes, fine-tuning stages, and evaluation metrics. Still, the comparison is useful because it shows where our simpler transfer learning pipeline stands relative to stronger recent systems.

Table 5: Comparison with recent related work on diabetic retinopathy classification and grading.

Study	Dataset	Method	Reported result
Chilukoti et al. [19]	EyePACS, APTOS, Messidor	Transfer learning and ensemble learning with QWK optimization	QWK 0.967 on APTOS
Shakibania et al. [20]	APTOS 2019 evaluation with multi-source training	Dual-branch model using ResNet50 and EfficientNetB0 feature extractors	89.60% accuracy and QWK 0.93 for stage grading
Vallukappully et al. [21]	APTOS 2019	NASNet-Large and ResNet50 with class balancing and preprocessing tests	Accuracy above 0.91 and F1-score 0.91 for multi-stage classification
Dharrao et al. [22]	APTOS 2019	EfficientNetB0 with multi-scale spatial attention gates	QWK 0.923 and macro AUC 0.976
Saadna et al. [23]	APTOS 2019 and Messidor-2	Lightweight transformer models with CLAHE preprocessing	QWK up to 0.89 and AUC up to 0.95
Al-Dolat et al. [24]	APTOS 2019 and DDR	CheXNet-CBAM with Grad-CAM visualization	96.12% accuracy on APTOS 2019
Huang and Huang [26]	APTOS 2019 and IDRiD	Position-wise attention-enhanced Vision Transformer	84% accuracy on APTOS 2019
Our project	APTOS 2019 with 80/20 validation split	Transfer learning with ResNet50 and EfficientNetB0, class weights, QWK, and Grad-CAM	69.03% accuracy, 0.5250 macro F1, and 0.8004 QWK for ResNet50

The comparison shows that our result is lower than many recent published systems, which is expected. The studies above do not all use the same split, training setup, preprocessing, architecture, or metric. Several of them use stronger preprocessing, full fine-tuning, attention modules, ensemble learning, transformer-based architectures, larger training settings, or external evaluation. Our contribution is more modest: we implemented a clear and reproducible transfer learning pipeline, compared two CNN backbones, reported macro metrics and QWK for the imbalanced five-class setting, and added Grad-CAM for visual inspection. This makes the project academically valid, but it also shows exactly what would need to be improved before aiming for a publishable model.

5.5 Class-Wise Interpretation

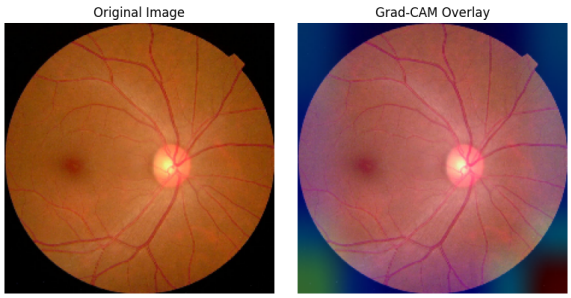
The class-wise results are important because the five DR classes are not equally represented. No DR was the easiest class for both models. This is probably because it has the largest number

of images and because normal retinal images are often easier to separate from advanced disease stages. Mild and Moderate were more difficult because their visual differences can be subtle. Severe was the weakest class in both models, with an F1-score close to 0.30. Proliferative DR also remained difficult, with F1-scores below 0.50 for both models.

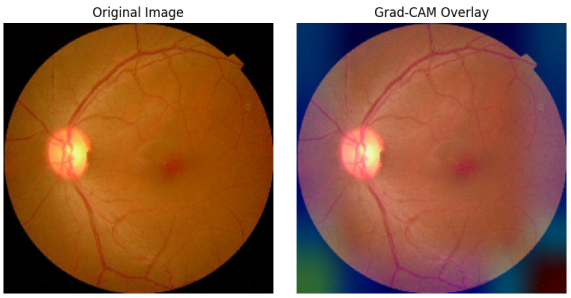
For a medical image task, these errors matter. A model that performs well mainly on No DR is not enough. From our results, transfer learning was helpful, but it also had clear limits on the smaller disease classes. ResNet50 gave the best overall summary result in this run, but the minority-class performance still needs work. A stronger project in the future could tune fine-tuning more systematically, use a larger dataset, apply more careful preprocessing, or use methods designed for ordinal classification.

6 Explainability with Grad-CAM

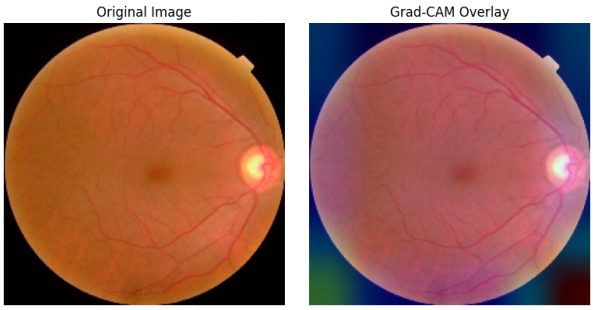
For a medical image classification project, it is useful to see what parts of an image influenced the model prediction. We used Grad-CAM to create heatmap overlays for selected validation images. The goal was not to prove clinical correctness, but to make the model behavior easier to inspect.



(a) No DR case.



(b) No DR case.



(c) No DR case.

Figure 7: Grad-CAM visualizations from validation samples. The examples show three No DR validation images predicted as No DR.

The Grad-CAM examples show the regions that contributed most to the prediction. This kind of visualization is helpful because a deep learning model should not be treated only as a black box, especially in healthcare tasks. The Grad-CAM examples were useful mainly as a visual check of the model behavior, not as clinical proof.

In the three examples, the model correctly predicted No DR cases. The heatmaps helped us inspect whether the model was reacting to retinal structures rather than only to the image border. These examples are limited because they show correct predictions from the same class, but

they still add a useful interpretation layer to the classification results. A stronger interpretation section in future work should include more classes and wrong predictions.

7 Discussion

The project shows that transfer learning is a practical approach for diabetic retinopathy grading. In the final notebook run, ResNet50 gave better validation performance than EfficientNetB0, while EfficientNetB0 remained faster and lighter to train. This result is useful because it shows that efficiency alone does not guarantee better validation performance on an imbalanced medical image dataset. At the same time, we do not want to overstate the results.

Compared with recent related work, our result is best understood as a baseline rather than a final state-of-the-art system. Recent studies on APTOS 2019 have reported higher performance using stronger strategies such as attention modules, transformer architectures, ensemble learning, extensive preprocessing, or more extensive fine-tuning [19, 20, 21, 22, 24]. This does not make our result useless; it makes the position of the project clearer. Our pipeline shows what can be achieved with a simpler transfer learning setup, and the gap with recent work points to the next steps needed if the project is developed further.

The model performs strongest on the No DR class because that class has the most images and is visually easier to separate from diseased classes. The Severe and Proliferative DR classes are harder. The confusion matrix shows that many Severe images are confused with Moderate or Proliferative DR. This was not surprising because these classes are close in disease progression, and the Severe class had only 39 validation images.

Another important point is that validation accuracy can look better than macro F1-score. Accuracy is influenced by the majority class, while macro F1 gives equal importance to all classes. For this reason, macro F1 is more useful for understanding the real performance across all disease grades. The project also shows why explainability is useful. Grad-CAM does not prove that the model is clinically correct, but it helps us inspect the prediction process. In a real medical system, this would need to be combined with expert review, external validation, and more detailed analysis of wrong predictions.

7.1 What the Results Mean for the Project

The results suggest that the pipeline learned useful visual patterns from the retinal images, especially for the larger classes. ResNet50 achieved higher accuracy, macro F1-score, and QWK than EfficientNetB0, so the final model comparison is supported by the notebook outputs. The model also performed especially well on the No DR class, which is expected because this class has the largest number of images and often has clearer visual separation from diseased classes.

The QWK result adds another useful view of the same experiment. ResNet50 achieved a QWK score of 0.8004, compared with 0.7931 for EfficientNetB0. This supports the ResNet50 result not only by standard classification metrics, but also by an ordinal grading metric that is commonly used in diabetic retinopathy studies.

However, the results also show that the task is not solved. The minority classes are still difficult. This is important because the advanced disease stages are clinically important, but they are also the classes with fewer samples in the dataset. In our opinion, this is one of the most important findings of the project. The model can classify many images correctly, but the class-wise results show where it is still weak.

7.2 Future Improvements

If we continue this project, the first improvement would be to add an independent test set. A validation split is useful during development, but a separate test set gives a stronger final evaluation. A second improvement would be to tune the fine-tuning stage more carefully. In this run, fine-tuning was included for both models, but EfficientNetB0 still did not outperform ResNet50. More controlled fine-tuning settings, such as different learning rates, fewer unfrozen layers, or early stopping based on macro F1/QWK, could be tested.

Another improvement would be to test methods that are designed for imbalanced or ordinal classification. For example, focal loss could give more attention to difficult samples, and ordinal regression or QWK-based optimization could use the ordered relationship between DR grades. This would match the way recent DR grading studies evaluate disagreement between nearby and distant disease stages. More Grad-CAM examples should also be added, especially wrong predictions. Looking at wrong predictions can help explain whether the model is confusing visually similar disease stages or reacting to image quality issues.

Finally, the project could be improved by involving medical feedback. A clinician or trained grader could review selected Grad-CAM heatmaps and confusion cases. This would not automatically make the model clinically valid, but it would make the interpretation stronger and more meaningful.

8 Limitations

This project has some limitations:

- The dataset is imbalanced, especially for the Severe class.
- The notebook uses a train-validation split, but an independent test set would give a stronger final evaluation.
- Fine-tuning was executed, but only with a limited setting. More careful hyperparameter tuning may change the final comparison.
- Grad-CAM examples were shown for only a few validation images. More examples, including wrong predictions, would make the interpretation stronger.
- The model was not evaluated in a real clinical workflow, so the results should be interpreted as course project results, not as clinical evidence.

9 Conclusion

In this project, we built a deep learning pipeline for diabetic retinopathy grading using transfer learning CNNs. ResNet50 and EfficientNetB0 were trained and compared using the same validation split. In the final notebook run, ResNet50 achieved the better validation performance, with 0.6903 accuracy, 0.5250 macro F1-score, and 0.8004 QWK. The results show that transfer learning can classify retinal images reasonably well, but minority classes remain challenging. Grad-CAM added an explainability component by showing which image areas influenced the predictions. Overall, the project demonstrates the use of pretrained CNNs, class weighting, fine-tuning, ordinal-aware evaluation, model comparison, and explainable AI for a medical image classification problem.

References

[1] 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, 2016.

[2] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in International Conference on Machine Learning, 2019.

[3] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in Proceedings of the IEEE International Conference on Computer Vision, 2017.

[4] V. Gulshan et al., "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," JAMA, vol. 316, no. 22, pp. 2402–2410, 2016.

[5] D. S. W. Ting et al., "Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes," JAMA, vol. 318, no. 22, pp. 2211–2223, 2017.

[6] R. Gargeya and T. Leng, "Automated Identification of Diabetic Retinopathy Using Deep Learning," Ophthalmology, vol. 124, no. 7, pp. 962–969, 2017.

[7] M. D. Abramoff et al., "Pivotal Trial of an Autonomous AI-Based Diagnostic System for Detection of Diabetic Retinopathy in Primary Care Offices," npj Digital Medicine, vol. 1, article 39, 2018.

[8] N. Tsiknakis et al., "Deep Learning for Diabetic Retinopathy Detection and Classification Based on Fundus Images: A Review," Computers in Biology and Medicine, vol. 135, article 104599, 2021.

[9] Z. Zhelev et al., "Test Accuracy of Artificial Intelligence-Based Grading of Fundus Images in Diabetic Retinopathy Screening: A Systematic Review," Journal of Medical Screening, vol. 30, no. 3, pp. 97–112, 2023.

[10] H. Naz, N. J. Ahuja, and R. Nijhawan, "Diabetic Retinopathy Detection Using Supervised and Unsupervised Deep Learning: A Review Study," Artificial Intelligence Review, vol. 57, article 131, 2024.

[11] N. Ul Haq et al., "Computationally Efficient Deep Learning Models for Diabetic Retinopathy Detection: A Systematic Literature Review," Artificial Intelligence Review, vol. 57, article 309, 2024.

[12] I. Abu-Shawish et al., "Deep Learning in Automatic Diabetic Retinopathy Detection and Grading Systems: A Comprehensive Survey and Comparison of Methods," IEEE Access, vol. 12, pp. 84785–84802, 2024.

[13] M. Vijayan and V. S., "A Regression-Based Approach to Diabetic Retinopathy Diagnosis Using EfficientNet," Diagnostics, vol. 13, no. 4, article 774, 2023.

[14] S. Batool et al., "Deploying Efficient Net Batch Normalizations for Grading Diabetic Retinopathy Severity Levels From Fundus Images," Scientific Reports, vol. 13, article 14462, 2023.

[15] M. N. Alam et al., "Contrastive Learning-Based Pretraining Improves Representation and Transferability of Diabetic Retinopathy Classification Models," Scientific Reports, vol. 13, article 6047, 2023.

[16] G. Alwakid, W. Gouda, M. Humayun, and N. Z. Jhanjhi, "Enhancing Diabetic Retinopathy Classification Using Deep Learning," Digital Health, vol. 9, 2023.

[17] A. Malhi, R. Grewal, and H. S. Pannu, "Detection and Diabetic Retinopathy Grading Using Digital Retinal Images," International Journal of Intelligent Robotics and Applications, vol. 7, pp. 426–458, 2023.

[18] D. Mane et al., "Diabetic Retinopathy Recognition and Classification Using Transfer Learning Deep Neural Networks," Traitement du Signal, vol. 41, no. 5, pp. 2683–2691, 2024.

[19] S. V. Chilukoti, L. Shan, V. S. Tida, A. S. Maida, and X. Hei, "A Reliable Diabetic Retinopathy Grading via Transfer Learning and Ensemble Learning with Quadratic Weighted Kappa Metric," BMC Medical Informatics and Decision Making, vol. 24, article 37, 2024.

[20] H. Shakibania, S. Raoufi, B. Pourafkham, H. Khotanlou, and M. Mansoorizadeh, "Dual Branch Deep Learning Network for Detection and Stage Grading of Diabetic Retinopathy," Biomedical Signal Processing and Control, vol. 93, article 106168, 2024.

[21] S. Vallukappully, I. van der Linde, and A. Chakraborty, "Early Detection and Classification of Diabetic Retinopathy by Transfer Learning of NASNet-Large and ResNet-50 Convolutional Neural Networks," Informatics in Medicine Unlocked, vol. 58, article 101688, 2025.

[22] D. Dharrao et al., "AI-Driven Detection and Classification of Diabetic Retinopathy Stages Using EfficientNetB0," Discover Applied Sciences, vol. 7, article 1400, 2025.

[23] Y. Saadna, S. Mezzoudj, and M. Khelifa, "Efficient Transformer Architectures for Diabetic Retinopathy Classification from Fundus Images: DR-MobileViT, DR-EfficientFormer, and DR-SwinTiny," Informatica, vol. 49, no. 29, pp. 117–128, 2025.

[24] W. Al-Dolat et al., "Enhancing Fundus Image Analysis for Diabetic Retinopathy Using CheXNet with CBAM and Grad-CAM Visualization," Frontiers in Medicine, vol. 13, article 1732109, 2026.

[25] J. Civit-Masot et al., "An Explainable Ensemble for Diabetic Retinopathy Grading with a Novel Confidence Quality Factor and Configurable Heatmaps," Medical & Biological Engineering & Computing, vol. 64, pp. 1231–1247, 2026.

[26] Y.-H. Huang and Y.-T. Huang, "Position-Wise Attention-Enhanced Vision Transformer for Diabetic Retinopathy Grading," Computers, Materials & Continua, vol. 87, no. 3, article 45, 2026.