

Impact of Class Imbalance on Deep Learning-Based Diabetic Retinopathy Classification: An Analysis of the APTOS 2019 Dataset

Santhana Sagaya Mary A^{1*}, Dr.S. Chidambaranathan²

¹Research Scholar, Reg. No. 23221282282009,
Department of Information Technology
PG and Research Department of Computer Science, herlinstar@gmail.com

²Associate Professor, Department of Computer Applications,
^{1&2} St. Xavier's College (Autonomous), Palayamkottai,
^{1&2} Affiliated to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli- 627 012,
Tamil Nadu, India. scharan2009@gmail.com

DOI: [https://doi.org/10.63001/tbs.2026.v21.i03.S.I\(3\).pp1986-1994](https://doi.org/10.63001/tbs.2026.v21.i03.S.I(3).pp1986-1994)

KEYWORDS

Keywords: Diabetic retinopathy, transfer learning, APTOS 2019, retinal fundus images, precision, recall, F1-score.

Article Info

Received : 10/05 /2026

Accepted : 15/06/2026

Published: 20/07/2026

Abstract

Diabetic retinopathy (DR) is a major microvascular complication of diabetes and an important cause of preventable visual impairment. The increasing use of deep learning for retinal image analysis has demonstrated promising diagnostic performance; however, the distribution of disease severity within publicly available datasets can substantially influence the interpretation of classification performance. This study analyses the implications of class imbalance in the APTOS 2019 diabetic retinopathy dataset and examines how imbalance should be considered when interpreting transfer-learning-based classification metrics. The APTOS 2019 dataset contains 3,662 retinal fundus images distributed across five severity categories: No Diabetic Retinopathy, Mild, Moderate, Severe and Proliferative Diabetic Retinopathy. The dataset exhibits considerable class imbalance, with 1,805 images in the No DR category and only 193 images in the Severe DR category, corresponding to a largest-to-smallest class ratio of 9.35:1. Reported performance values for DenseNet121, ResNet50, VGG16, InceptionV3 and MobileNetV2 were examined using accuracy, area under the curve (AUC), precision, recall and F1-score. Although the reported accuracy values were generally high, substantial differences were observed between precision and recall for some architectures. MobileNetV2, for example, achieved 99.64% precision but only 74.57% recall and an F1-score of 85.22%, whereas VGG16 demonstrated a more balanced precision of 94.68% and recall of 95.42%. These findings highlight the importance of considering multiple performance indicators rather than relying on accuracy alone when evaluating multiclass DR classification under imbalanced class distributions. The analysis provides a dataset-level perspective on class imbalance and its implications for responsible interpretation of deep-learning-based diabetic retinopathy classification.

1. Introduction

Diabetic retinopathy (DR) is one of the most important microvascular complications associated

with diabetes mellitus and remains a major cause of visual impairment among working-age populations. Early identification of retinal abnormalities is essential because timely referral and treatment can reduce the risk of progression to advanced stages of

visual loss. Conventional DR assessment is based on examination of retinal fundus photographs by trained ophthalmologists or graders. However, the increasing volume of retinal images and the limited availability of specialist expertise have encouraged the development of automated computer-aided screening and classification systems.

Deep learning (DL), particularly convolutional neural networks (CNNs), has emerged as an important approach for automated retinal image analysis. Recent systematic evidence has demonstrated the potential of deep-learning methods for detecting diabetic retinopathy from retinal images and optical coherence tomography. Bi *et al.* (2025), in a systematic review and meta-analysis involving 47 studies and 188,268 retinal images and OCT scans, reported strong diagnostic potential of deep-learning approaches while emphasizing the importance of dataset standardization, model interpretability and validation across diverse populations.

Recent evidence also indicates that the performance of artificial intelligence systems cannot be interpreted independently of the characteristics of the dataset on which they are evaluated. Wang *et al.* (2026) reviewed 82 studies involving 887,244 examinations across 25 regulator-approved deep-learning systems and found that performance varied according to disease-severity threshold, image gradability, diagnostic criteria, reference standard and other contextual factors. Their findings emphasize that apparently high diagnostic performance does not necessarily imply uniform performance across different clinical and imaging conditions.

One important methodological issue in multiclass

DR classification is class imbalance. In a balanced dataset, different disease categories are represented in approximately comparable numbers. In contrast, an imbalanced dataset contains substantially more observations in some categories than in others. This is particularly relevant to DR because severe and proliferative stages generally occur less frequently than normal or early-stage retinal findings. When one class dominates the dataset, overall accuracy can provide an incomplete representation of classification behaviour, particularly when the clinical importance of minority classes is considered. The APTOS 2019 dataset is widely used for automated DR classification and contains 3,662 retinal fundus images distributed across five categories. The distribution is highly uneven: 1,805 images (49.29%) belong to the No DR category, whereas only 193 images (5.27%) belong to the Severe DR category. Thus, the largest class is approximately 9.35 times larger than the smallest class.

Recent research has specifically highlighted the need to address class imbalance in DR classification. Soomro *et al.* (2025), for example, investigated an approach combining CNN-based feature extraction with SMOTE and data augmentation for imbalanced DR classification. Their study reported improved classification performance through explicit treatment of class imbalance, demonstrating the continuing methodological importance of this issue in automated retinal image analysis.

The problem is not limited to dataset composition. Class imbalance also influences how performance metrics should be interpreted. Accuracy provides an overall proportion of correctly classified observations but does not reveal whether minority

disease categories are being adequately recognized. Precision indicates the proportion of predicted positive cases that are correct, whereas recall reflects the ability to identify actual positive cases. F1-score provides a harmonic balance between precision and recall. Consequently, differences between these measures can provide important information that is not visible through accuracy alone.

This issue is particularly relevant in the interpretation of published transfer-learning results for APTOS 2019. Five commonly used architectures—DenseNet121, ResNet50, VGG16, InceptionV3 and MobileNetV2—have reported high overall performance, but their precision, recall and F1-score values are not uniformly distributed. For example, MobileNetV2 reported 99.64% precision but only 74.57% recall and an F1-score of 85.22%, whereas VGG16 reported 94.68% precision and 95.42% recall with an F1-score of 95.02%.

Recent real-world evidence further reinforces the importance of examining performance beyond a single metric. Cheung *et al.* (2026) systematically evaluated prospective real-world validation and economic evidence for deep-learning-based DR detection and emphasized the need to understand performance in practical deployment environments rather than relying exclusively on controlled or retrospective evaluations.

Therefore, the present analysis focuses specifically on class imbalance and its implications for performance interpretation in the APTOS 2019 dataset. Rather than presenting another architecture-ranking study, the analysis examines the relationship between the dataset's class distribution and the behaviour of commonly reported evaluation metrics. Particular attention is given to the importance of

precision, recall and F1-score when interpreting deep-learning-based multiclass DR classification.

2. Materials and Methods

2.1 Study Design and Data Source

The present work employed a secondary-data analytical design based on the publicly available APTOS 2019 diabetic retinopathy dataset and reported performance results from transfer-learning-based classification studies.

The APTOS 2019 dataset consists of 3,662 retinal fundus images categorized into five diabetic retinopathy severity levels:

1. No Diabetic Retinopathy
2. Mild Diabetic Retinopathy
3. Moderate Diabetic Retinopathy
4. Severe Diabetic Retinopathy
5. Proliferative Diabetic Retinopathy

The dataset distribution was examined using class frequency and percentage. The degree of imbalance was quantified using the ratio between the largest and smallest classes. The dataset contains 1,805 No DR images and 193 Severe DR images, resulting in a largest-to-smallest class ratio of **9.35:1**.

No new patients were recruited for the present analysis. No clinical examinations, retinal image acquisition or patient-level interventions were performed.

2.2 Transfer-Learning Architectures and Performance Measures

Five widely used pretrained CNN architectures were considered:

- DenseNet121
- ResNet50

- VGG16
- InceptionV3
- MobileNetV2

These architectures were selected because they represent different CNN design strategies and have been extensively applied to medical image classification. The architecture descriptions considered in the analysis are documented in the existing study material.

The following performance measures were examined:

Accuracy:

The proportion of correctly classified observations among all observations.

AUC:

The area under the receiver operating characteristic curve, representing the ability of a classifier to discriminate between categories across classification thresholds.

Precision:

The proportion of predicted positive classifications that are actually positive.

Recall:

The proportion of actual positive cases correctly identified by the classifier.

F1-score:

The harmonic mean of precision and recall.

The reported model performance values were treated as secondary/literature-reported results rather than independently reproduced experimental results. Therefore, no claim is made that the models were retrained or independently validated in the present investigation.

2.3 Analytical Approach

The analysis proceeded in three stages. First, the

distribution of the five DR categories was quantified to identify the extent of class imbalance. Second, reported transfer-learning performance was examined across accuracy, AUC, precision, recall and F1-score. Third, differences between precision and recall were considered to determine whether high overall performance was accompanied by balanced metric behaviour.

The analysis was deliberately focused on metric interpretation under an imbalanced class distribution rather than on establishing causal relationships between class imbalance and individual model performance.

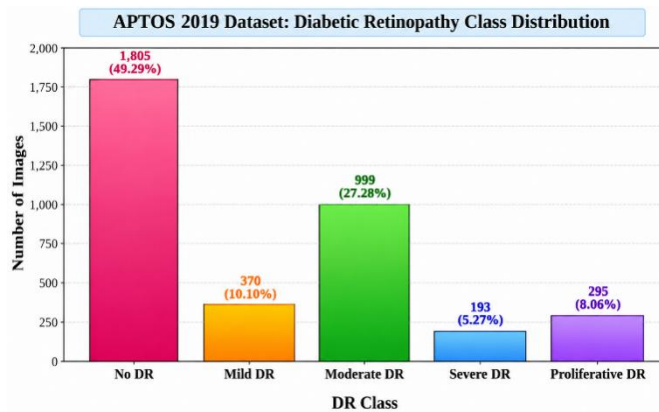
3. Results

3.1 Distribution of the APTOS 2019 Dataset

The APTOS 2019 dataset contained 3,662 retinal fundus images distributed across five DR severity categories. No DR represented the largest category, with 1,805 images (49.29%), whereas Severe DR represented the smallest category, with 193 images (5.27%). Moderate DR accounted for 999 images (27.28%), followed by Proliferative DR with 295 images (8.06%) and Mild DR with 370 images (10.10%).

Table 1. Distribution of APTOS 2019 Dataset and Class-Imbalance Ratio

DR Class	Number of Images	Percentage (%)
No DR	1,805	49.29
Mild DR	370	10.10
Moderate DR	999	27.28
Severe DR	193	5.27
Proliferative DR	295	8.06
Total	3,662	100.00



The difference between the largest and smallest categories was substantial. The No DR category was approximately 9.35 times larger than the Severe DR category. This indicates a pronounced imbalance across the five classes.

The distribution can therefore be characterized by a dominant non-DR category, a comparatively large Moderate DR category and relatively small Severe and Proliferative DR categories. Such a distribution is important when interpreting aggregate classification metrics because overall performance may be strongly influenced by the more frequently represented categories.

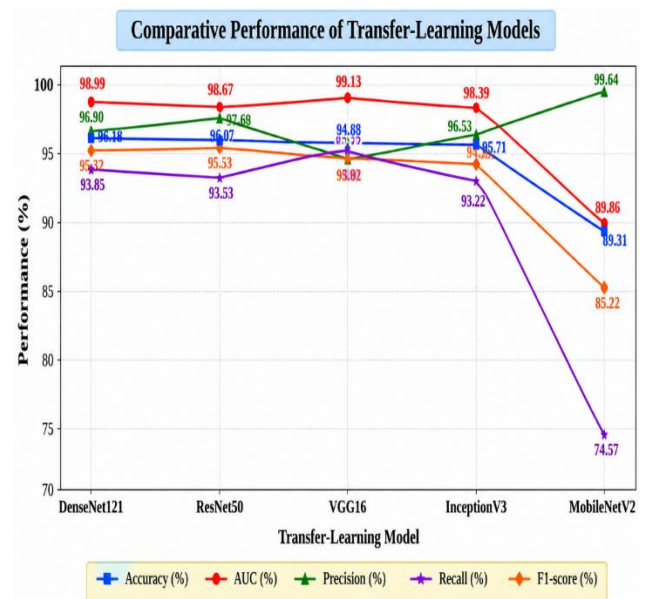
3.2 Comparative Performance of Transfer-Learning Models

The reported performance of the five transfer-learning architectures demonstrated relatively high accuracy and AUC for four of the five models, while MobileNetV2 showed substantially lower performance across several measures. The reported values are summarized below.

Table 2. Reported Performance of Transfer-Learning Models

Model	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)	F1-score (%)
DenseNet1	96.18	98.9	96.90	93.85	95.3

21		9			2
ResNet50	96.07	98.6	97.69	93.53	95.5
		7			3
VGG16	95.85	99.1	94.68	95.42	95.0
		3			2
InceptionV	95.71	98.3	96.53	93.22	94.8
3		9			1
MobileNet	89.31	89.8	99.64	74.57	85.2
V2		6			2



The reported accuracy values ranged from 89.31% to 96.18%, representing a difference of 6.87 percentage points. DenseNet121 recorded the highest reported accuracy at 96.18%, while MobileNetV2 recorded the lowest at 89.31%.

A similar pattern was observed for AUC. VGG16 achieved the highest reported AUC of 99.13%, followed by DenseNet121 at 98.99%, ResNet50 at 98.67% and InceptionV3 at 98.39%. MobileNetV2 recorded an AUC of 89.86%.

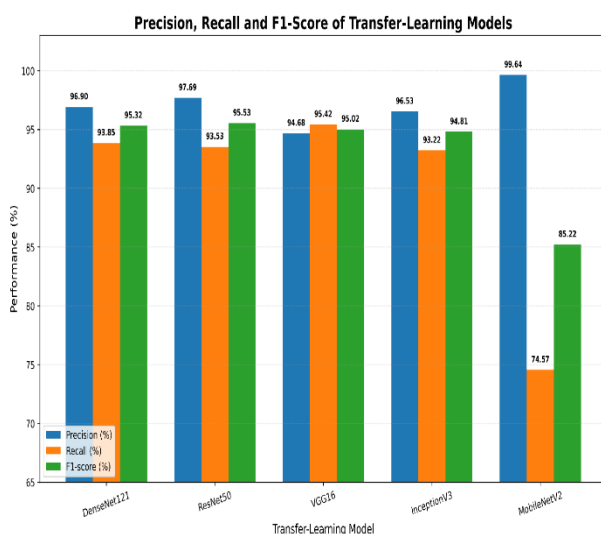
3.3 Precision, Recall and F1-Score Behaviour

The relationship between precision and recall revealed an important difference that was not apparent from accuracy alone. DenseNet121, ResNet50, VGG16 and InceptionV3 demonstrated

relatively close precision and recall values. In contrast, MobileNetV2 showed a substantial divergence between the two measures.

Table 3. Relationship Between Precision, Recall and F1-Score

Model	Precision (%)	Recall (%)	F1-score (%)	Interpretation
DenseNet121	96.90	93.85	95.32	Relatively balanced performance
ResNet50	97.69	93.53	95.53	High precision with strong recall
VGG16	94.68	95.42	95.02	Highly balanced precision and recall
InceptionV3	96.53	93.22	94.81	Relatively balanced performance
MobileNetV2	99.64	74.57	85.22	Very high precision but substantially lower recall



MobileNetV2 exhibited the largest precision–recall difference, with a difference of 25.07 percentage points. In comparison, VGG16 showed only a 0.74 percentage-point difference between precision and recall.

This finding is particularly relevant to an imbalanced multiclass classification problem. A model may achieve very high precision while failing to identify a considerable proportion of actual cases. Therefore, the precision value of 99.64% for MobileNetV2 should not be interpreted independently of its recall of 74.57%.

The F1-score provides a more integrated representation of this relationship. ResNet50 achieved the highest reported F1-score at 95.53%, followed by DenseNet121 at 95.32%, VGG16 at 95.02% and InceptionV3 at 94.81%. MobileNetV2 had a considerably lower F1-score of 85.22%.

4. Discussion

The present analysis demonstrates that class imbalance is an important consideration when interpreting deep-learning-based diabetic retinopathy classification results. The APTOS 2019 dataset contains a marked difference between the most and least represented classes, with the No DR category accounting for almost half of the dataset and the Severe DR category representing only 5.27%. The largest-to-smallest class ratio of 9.35:1 indicates that the five categories are not equally represented.

The observed imbalance is particularly important in a five-class disease-severity classification problem. A model evaluated predominantly on the more frequent categories may demonstrate high overall accuracy without necessarily demonstrating equivalent recognition capability across all disease

stages. Therefore, accuracy should be interpreted together with class-sensitive measures such as precision, recall and F1-score.

The reported performance results provide a useful illustration of this issue. DenseNet121, ResNet50, VGG16 and InceptionV3 produced accuracy values above 95%, and their precision and recall values were relatively close. In contrast, MobileNetV2 demonstrated a distinctive metric pattern: its precision was 99.64%, but its recall was only 74.57%. Consequently, its F1-score decreased to 85.22%.

The MobileNetV2 result should not be interpreted as proof that class imbalance alone caused the lower recall. Model architecture, training strategy, preprocessing, threshold selection, data partitioning and other methodological factors may also contribute to differences in performance. However, the result clearly demonstrates why a single aggregate metric is insufficient for evaluating an imbalanced multiclass medical-image classification problem.

The contrast between MobileNetV2 and VGG16 is particularly informative. VGG16 showed precision of 94.68% and recall of 95.42%, producing an F1-score of 95.02%. The small difference of only 0.74 percentage points suggests comparatively balanced metric behaviour. MobileNetV2, on the other hand, had a 25.07 percentage-point difference between precision and recall. Thus, although MobileNetV2 had the highest precision among the five architectures, this did not translate into similarly high recall or F1-score.

Recent literature supports the need for explicit consideration of imbalance. Soomro et al. (2025) specifically addressed imbalanced DR classification

through SMOTE and data augmentation, reporting 92.73% multiclass accuracy. Their work illustrates that class distribution is not merely a descriptive characteristic of a dataset but an important methodological consideration during model development and evaluation.

Similarly, recent systematic reviews emphasize that deep-learning performance varies according to dataset and evaluation context. Bi *et al.* (2025) highlighted the need for standardized datasets, interpretability and validation across diverse populations. Wang *et al.* (2026) further demonstrated substantial heterogeneity among regulator-approved DR systems and identified disease-severity thresholds, image gradability, diagnostic criteria and reference standards as important sources of performance variation.

These observations have direct implications for DR classification research. If minority classes such as Severe and Proliferative DR are clinically important but comparatively underrepresented, researchers should report class-wise performance wherever possible. Macro-averaged metrics, confusion matrices, sensitivity/recall for individual severity categories and F1-score can provide a more informative assessment than overall accuracy alone. The findings also have implications for the comparison of transfer-learning architectures. Ranking models solely according to accuracy can obscure differences in classification behaviour. For example, DenseNet121 had the highest reported accuracy, whereas ResNet50 had the highest reported F1-score, and VGG16 had the highest reported AUC. MobileNetV2 had the highest precision but the lowest recall and F1-score.

This perspective is also relevant to clinical

translation. Recent real-world evidence indicates that prospective validation of deep-learning DR systems remains important when moving from retrospective datasets to practical screening environments. Therefore, performance reported on a public dataset should not automatically be assumed to represent equivalent performance in clinical practice.

From these results suggest that class imbalance should be treated as an important methodological characteristic of APTOS 2019 when interpreting transfer-learning-based DR classification. The findings do not establish a causal effect of imbalance on any individual architecture; rather, they demonstrate the interpretive implications of an uneven class distribution and the importance of multi-metric evaluation.

5. Conclusion

The present analysis demonstrates that class imbalance is an important consideration in the interpretation of deep-learning-based diabetic retinopathy classification using the APTOS 2019 dataset. The dataset contains 3,662 retinal fundus images across five severity categories, with a substantial difference between the largest and smallest classes. The 9.35:1 largest-to-smallest class ratio illustrates the uneven representation of disease categories.

Although high classification accuracy was reported for several transfer-learning architectures, comparison of precision, recall and F1-score revealed important differences in metric behaviour. The MobileNetV2 result, with 99.64% precision but only 74.57% recall and 85.22% F1-score, demonstrates why high precision or accuracy alone should not be interpreted as evidence of uniformly

strong classification performance.

The findings emphasize the importance of multi-metric evaluation, class-wise assessment and explicit reporting of dataset imbalance in diabetic retinopathy classification research. Future investigations should incorporate balanced or imbalance-aware training strategies, class-wise evaluation and independent external validation to establish the robustness and clinical reliability of deep-learning-based DR classification systems.

References

1. Bi, Z., Li, J., Liu, Q., & Fang, Z. (2025). Deep learning-based optical coherence tomography and retinal images for detection of diabetic retinopathy: a systematic and meta analysis. *Frontiers in Endocrinology*, 16, 1485311. <https://doi.org/10.3389/fendo.2025.1485311>
1. [Frontiers](#)
2. Soomro, D. B., ChengLiang, W., Ashraf, M., et al. (2025). Automated dual CNN-based feature extraction with SMOTE for imbalanced diabetic retinopathy classification. *Image and Vision Computing*, 105537. <https://doi.org/10.1016/j.imavis.2025.105537>
7. [ScienceDirect](#)
3. Wang, T. W., Luo, W. T., Tu, Y. K., Chou, Y. B., & Wu, Y. T. (2026). Systematic review and meta-analysis of regulator-approved deep learning systems for fundus diabetic retinopathy detections. *npj Digital Medicine*, 9, 110. <https://doi.org/10.1038/s41746-025-02223-7>
8. [Nature](#)

4. Cheung, C. Y., et al. (2026). Real-World Prospective Validation and Economic Evaluation of Deep Learning-Based Diabetic Retinopathy Detection From Fundus Photographs: A Systematic Review and Meta-analysis. *Diabetes Care*, 49(3), 510–525. <https://doi.org/10.2337/dc25-1493>. *Diabetes Journals*
5. APTOS 2019 Blindness Detection Dataset. Kaggle diabetic retinopathy fundus image dataset.
6. Existing reported transfer-learning analysis of APTOS 2019. Performance values for DenseNet121, ResNet50, VGG16, InceptionV3 and MobileNetV2 as documented in the present thesis analysis.