

Comparative Analysis of Transfer-Learning Models for Five-Class Diabetic Retinopathy Classification: A Secondary-Data Study Using 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.i04.S.I\(4\).pp545-554](https://doi.org/10.63001/tbs.2026.v21.i04.S.I(4).pp545-554)

KEYWORDS

Diabetic retinopathy,
retinal fundus images,
APTOS 2019, transfer
learning, deep learning

Article Info

Received : 18/08 /2026

Accepted : 22 09/2026

Published: 21/10/2026

Abstract

Diabetes-related retinal complications are a major cause of preventable visual impairment, with diabetic retinopathy being one of the most significant ocular complications. Recent advances in deep learning and transfer-learning techniques have demonstrated considerable potential for the automated detection and classification of diabetic retinopathy using retinal fundus images. The present study employed a secondary-data-based comparative approach using the publicly available APTOS 2019 dataset, comprising 3,662 images across five DR severity classes. Class distribution and imbalance were analysed, and published performance of commonly used transfer-learning models was compared using accuracy, AUC, precision, recall and F1-score. The dataset showed substantial class imbalance, with No DR accounting for 49.29% of images and Severe DR for 5.27%, resulting in a 9.35:1 largest-to-smallest class ratio. Published results from an APTOS 2019 evaluation reported accuracies of 96.18% for DenseNet121, 96.07% for ResNet50, 95.85% for VGG16, 95.71% for InceptionV3 and 89.31% for MobileNetV2. The findings highlight the importance of considering class distribution and multiple performance measures when comparing transfer-learning models for multi-class DR classification. This secondary-data analysis provides a basis for subsequent computational validation using the same public dataset.

1. Introduction

Diabetic retinopathy (DR) is a progressive retinal disorder associated with diabetes mellitus and represents an important cause of visual impairment worldwide. The disease develops through pathological changes in the retinal

microvasculature and may progress from mild non-proliferative abnormalities to severe disease and proliferative diabetic retinopathy. Early detection and appropriate grading are therefore important for preventing or delaying vision loss and for supporting timely clinical management (Raman *et al.*, 2019; Tsiknakis *et*

al., 2021).

Retinal fundus photography is widely used for documenting retinal abnormalities associated with diabetic retinopathy. Conventional interpretation of fundus photographs depends on trained ophthalmologists or graders who evaluate retinal lesions and assign a severity category. Although expert assessment remains important, the increasing volume of retinal images generated through screening programmes has encouraged the development of computer-aided methods for automated or assisted classification (Raman *et al.*, 2019; Bhulakshmi & Rajput, 2024).

Traditional computer-aided approaches to DR analysis commonly relied on manually designed image features, including colour, texture, vascular characteristics, microaneurysms, haemorrhages and other retinal structures. Machine-learning algorithms were subsequently applied to these extracted features. However, such approaches may be influenced by image quality, illumination variation, feature-engineering requirements and differences between datasets (Ting *et al.*, 2017; Tsiknakis *et al.*, 2021).

Deep learning has substantially changed this approach by enabling feature extraction and classification within a unified neural-network framework. Convolutional neural networks (CNNs) can automatically learn hierarchical visual representations directly from retinal photographs and have consequently become

widely investigated for automated detection and grading of diabetic retinopathy (Gulshan *et al.*, 2016; Tsiknakis *et al.*, 2021).

Transfer learning has become particularly useful in medical image analysis because medical datasets are often smaller than the datasets required for training deep architectures entirely from the beginning. In transfer learning, a model pretrained on a large image dataset can be adapted to a specific target classification task. Several pretrained architectures, including VGG16, ResNet50, InceptionV3, DenseNet121 and MobileNetV2, have therefore been investigated for retinal image classification (Tan *et al.*, 2021; Bhulakshmi & Rajput, 2024).

The APTOS 2019 Blindness Detection dataset is a publicly available dataset widely used for five-class diabetic retinopathy classification. It contains 3,662 labelled retinal fundus images distributed across five severity categories: No DR, Mild DR, Moderate DR, Severe DR and Proliferative DR. The distribution is highly uneven, with 1,805 images in the No DR category and only 193 images in the Severe DR category (Pratt *et al.*, 2016; Duan *et al.*, 2022).

Such class imbalance is an important methodological consideration in automated DR classification. A model may obtain relatively high overall accuracy while showing poorer recognition of minority classes. Therefore, evaluation of multi-class DR classification should consider complementary measures such as precision, recall, F1-score and AUC rather

than relying exclusively on overall accuracy (Tsiknakis *et al.*, 2021; Bhulakshmi & Rajput, 2024).

The present study adopts a secondary-data-based comparative approach. The analysis uses the publicly documented characteristics of the APTOS 2019 dataset and published performance results of transfer-learning models. The objective is to examine the distributional characteristics of the dataset and compare reported model performance using commonly employed evaluation measures. This approach provides a structured basis for understanding the influence of class distribution and model architecture before subsequent computational validation using the same public dataset.

2. Materials and Methods

2.1 Research Design

The present investigation follows a **secondary-data comparative research design**. The analysis is based on information already available through a publicly accessible retinal image dataset and published scientific studies.

The study consists of two major components:

1. Descriptive analysis of the APTOS 2019 dataset.
2. Comparative analysis of published transfer-learning performance results obtained using the APTOS 2019 dataset.

2.2 Data Source

The APTOS 2019 Blindness Detection dataset was selected as the principal dataset because it

provides labelled retinal fundus photographs covering five clinically meaningful DR severity categories.

The dataset contains 3,662 retinal fundus images distributed across the following categories:

- Grade 0 – No DR
- Grade 1 – Mild DR
- Grade 2 – Moderate DR
- Grade 3 – Severe DR
- Grade 4 – Proliferative DR

The published dataset distribution is shown in Table 1.

Table 1. Distribution of APTOS 2019 Retinal Fundus Images

DR Grade	Severity Category	Number of Images	Percentage (%)
Grade 0	No DR	1,805	49.29
Grade 1	Mild DR	370	10.10
Grade 2	Moderate DR	999	27.28
Grade 3	Severe DR	193	5.27
Grade 4	Proliferative DR	295	8.06
Total	Five classes	3,662	100.00

The class counts and the five-class structure are consistently reported in studies using APTOS 2019.

2.3 Calculation of Class Distribution

The percentage representation of each class was calculated using:

$$\text{Class Percentage} = \frac{\text{Number of images in class}}{\text{Total number of images}} \times 100$$

For example, the percentage of No DR images was:

$$\frac{1805}{3662} \times 100 = 49.29\%$$

Similarly, the percentage of Severe DR images was:

$$\frac{193}{3662} \times 100 = 5.27\%$$

These calculations demonstrate the unequal representation of the five severity categories.

2.4 Analysis of Class Imbalance

Class imbalance was evaluated by comparing the largest and smallest classes.

The largest class was No DR:

$$N_{max} = 1805$$

The smallest class was Severe DR:

$$N_{min} = 193$$

Therefore:

$$\text{Imbalance Ratio} = \frac{1805}{193} = 9.35$$

Thus, the largest class contains approximately **9.35 times as many images as the smallest class.**

Table 2. Descriptive Class-Imbalance Analysis

Parameter	Value
Total images	3,662
Number of classes	5
Largest class	No DR
Largest class size	1,805
Smallest class	Severe DR

Smallest class size	193
Largest/smallest ratio	9.35:1
Largest class proportion	49.29%
Smallest class proportion	5.27%

This imbalance is important when interpreting classification accuracy because the majority class contributes substantially more observations to the overall metric.

2.5 Transfer-Learning Architectures Considered

Transfer-learning models commonly investigated for DR classification include:

VGG16

VGG16 uses a sequence of convolutional layers followed by fully connected classification layers. Its relatively simple and structured architecture has made it a frequently used baseline for medical image classification.

ResNet50

ResNet50 introduces residual connections that facilitate the training of deeper neural networks. Residual learning allows information to bypass intermediate layers through shortcut connections.

InceptionV3

InceptionV3 employs multiple convolutional operations with different receptive-field characteristics within its Inception modules. This allows the network to capture features at multiple spatial scales.

DenseNet121

DenseNet121 connects layers densely so that

features generated by earlier layers can be reused by subsequent layers. This promotes feature propagation and can reduce redundant feature learning.

MobileNetV2

MobileNetV2 is designed as a relatively lightweight CNN using depthwise separable convolutions and inverted residual structures. Its lower computational requirements make it relevant to resource-constrained applications.

These architectures have been repeatedly investigated for automated DR classification, although their reported performance depends strongly on the experimental protocol.

2.6 Evaluation Measures

The comparative literature commonly reports several classification metrics.

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy represents the proportion of correctly classified observations.

Precision

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

Precision indicates the proportion of positive predictions that are correct.

Recall/Sensitivity

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

Recall represents the ability of the model to identify relevant positive cases.

F1-score

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

F1-score provides a combined measure of precision and recall.

Area Under the ROC Curve

AUC summarizes discriminatory ability across classification thresholds. In multi-class classification, the calculation may be performed using one-vs-rest or other aggregation strategies depending on the study protocol.

Because the present work compares published results, the reported metric definitions and evaluation protocols were retained as described by the respective studies rather than recalculating unavailable patient-level predictions.

3. Results

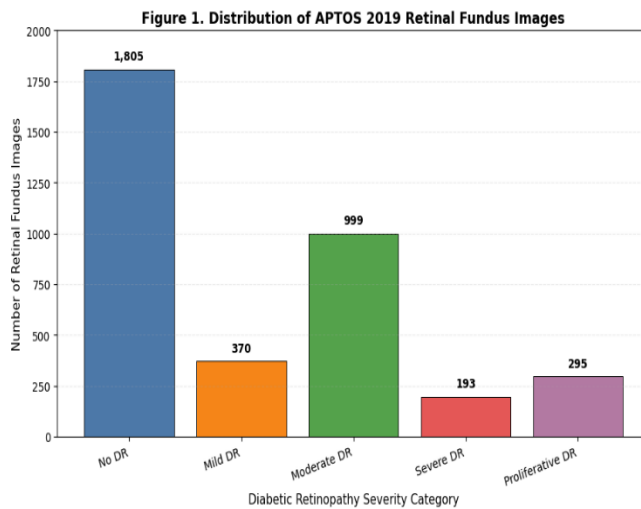
3.1 Distribution of the APTOS 2019 Dataset

The descriptive analysis revealed a substantial difference between the five severity categories. No DR constituted the largest group with 1,805 images, accounting for 49.29% of the dataset. Moderate DR was the second-largest category with 999 images (27.28%). Mild DR accounted for 370 images (10.10%), while Proliferative DR represented 295 images (8.06%). Severe DR was the smallest group, containing 193 images (5.27%).

The distribution therefore demonstrates that almost half of the dataset belongs to the No DR category, whereas the Severe DR category represents only slightly more than five percent of the available images.

Figure 1. Distribution of APTOS 2019 retinal

fundus images across five diabetic retinopathy severity categories.



3.2 Comparative Performance of Transfer-Learning Models

A published comparative investigation evaluated several pretrained CNN architectures using the APTOS 2019 dataset. The reported results included DenseNet121, MobileNetV2, ResNet50, InceptionV3 and VGG16. The comparison included accuracy, AUC, F1-score, precision and recall.

Table 3. Published Comparative Performance of Transfer-Learning Models on APTOS 2019

Model	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)	F1-score (%)
DenseNet121	96.18	98.99	96.90	93.85	95.32
MobileNetV2	89.31	89.86	99.64	74.57	85.22
ResNet50	96.07	98.67	97.69	93.53	95.53

Inception V3	95.71	98.39	96.53	93.22	94.81
VGG16	95.85	99.13	94.68	95.42	95.02

3.3 Accuracy Comparison

The reported accuracy values ranged from 89.31% to 96.18%.

The accuracy values were:

- DenseNet121 – 96.18%
- ResNet50 – 96.07%
- VGG16 – 95.85%
- InceptionV3 – 95.71%
- MobileNetV2 – 89.31%

The arithmetic mean of the five reported accuracy values was:

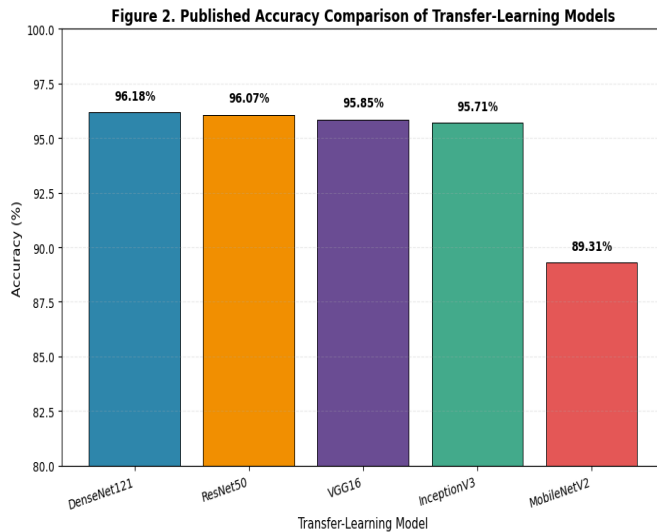
$$\frac{96.18 + 89.31 + 96.07 + 95.71 + 95.85}{5} = 94.62\%$$

The difference between the highest and lowest reported accuracy was:

$$96.18 - 89.31 = 6.87$$

Thus, the reported accuracy values varied across architectures by **6.87 percentage points** within this particular comparative study.

Figure 2. Published accuracy comparison of transfer-learning architectures on the APTOS 2019 dataset.



3.4 AUC Comparison

The reported AUC values ranged from 89.86% for MobileNetV2 to 99.13% for VGG16.

The difference between the highest and lowest AUC was:

$$99.13 - 89.86 = 9.27$$

This indicates that model discrimination can differ considerably even when the models are evaluated on the same underlying dataset.

3.5 Precision and Recall Comparison

The published precision values ranged from 94.68% to 99.64%.

MobileNetV2 demonstrated a relatively high reported precision of 99.64%, but its reported recall was substantially lower at 74.57%. This illustrates why precision should not be interpreted independently.

The difference between its reported precision and recall was:

$$99.64 - 74.57 = 25.07$$

By comparison, VGG16 had a reported precision of 94.68% and recall of 95.42%,

representing a much smaller difference between the two measures.

This observation demonstrates that a high precision value does not necessarily indicate balanced classification performance.

3.6 F1-Score Comparison

F1-score values ranged from 85.22% for MobileNetV2 to 95.53% for ResNet50 in the cited comparative evaluation.

The range was:

$$95.53 - 85.22 = 10.31$$

The F1-score therefore provides additional information beyond accuracy by incorporating both precision and recall.

4. Discussion

The present secondary-data analysis demonstrates two important characteristics of automated DR classification: substantial class imbalance within the dataset and variation in reported performance among transfer-learning architectures.

The APTOS 2019 dataset contains 3,662 images distributed across five DR severity categories. The No DR class represents approximately 49.29% of the complete dataset, whereas Severe DR represents approximately 5.27%. Previous studies using the same dataset have similarly emphasized the unequal distribution of the five classes.

The calculated 9.35:1 ratio between the largest and smallest classes indicates that the dataset cannot be considered balanced. This characteristic is particularly relevant for multi-

class DR classification because the minority categories are clinically important despite containing fewer observations.

The literature comparison also demonstrates that accuracy alone does not adequately characterize model behaviour. The reported accuracy values were relatively high for four of the five architectures, but the difference in recall was more substantial. MobileNetV2, for example, showed high reported precision but considerably lower recall in the cited study.

This difference has methodological significance. A model can produce a high overall accuracy or precision while failing to identify a proportion of relevant cases. For a medical classification problem, sensitivity/recall and F1-score therefore provide useful complementary information.

The results also illustrate that architecture names alone cannot determine expected performance. Transfer learning is influenced by multiple factors, including preprocessing, image quality, data augmentation, class balancing, training configuration, feature-extraction strategy and evaluation protocol. Consequently, published accuracy values from different studies should not automatically be treated as directly interchangeable.

Another important consideration is that different researchers may use different training-validation-test splits. For example, published studies using APTOS 2019 have employed different partitioning strategies, including

75:15:10 and 80:20 arrangements. Therefore, direct comparison across unrelated publications should be undertaken cautiously.

The secondary-data approach used in the present study provides a useful foundation for the planned thesis work. In the subsequent computational phase, the same APTOS 2019 dataset can be processed under a clearly defined experimental protocol. Models can then be trained and evaluated using identical preprocessing, data partitioning and performance measures. Such a controlled comparison would reduce methodological variation and allow original computational results to be distinguished clearly from previously published results.

The current analysis therefore serves two purposes. First, it establishes the statistical characteristics of the dataset. Second, it identifies the major evaluation dimensions that should be retained when conducting an original transfer-learning experiment.

5. Limitations

Several limitations should be considered when interpreting the present study.

First, the study is based on secondary information and publicly available dataset characteristics rather than newly collected clinical images.

Second, the comparative model values presented in Table 3 are reported results from a published study. They are not original predictions generated by the present

investigation.

Third, differences in preprocessing, image augmentation, training configuration, class balancing and dataset partitioning can influence reported performance. Consequently, literature-derived accuracy values should not be interpreted as a controlled head-to-head experiment unless the original studies used comparable protocols.

Fourth, the APTOS 2019 dataset itself is imbalanced. Minority classes, particularly Severe DR, contain substantially fewer images than the No DR class. This may affect the interpretation of overall accuracy.

Finally, future computational validation using the same dataset is required before making conclusions regarding the relative performance of architectures under a single standardized experimental protocol.

6. Conclusion

The present study provides a secondary-data comparative analysis of transfer-learning approaches for five-class diabetic retinopathy classification using the APTOS 2019 dataset.

The dataset contains 3,662 labelled retinal fundus images distributed across five DR severity categories. The descriptive analysis revealed considerable class imbalance, with No DR representing 49.29% of the dataset and Severe DR representing 5.27%. The largest-to-smallest class ratio was approximately 9.35:1.

Published comparative results further demonstrated differences among commonly

used transfer-learning architectures. DenseNet121, ResNet50, VGG16 and InceptionV3 produced reported accuracy values above 95% in the selected comparative evaluation, whereas MobileNetV2 showed a lower reported accuracy in that particular study. However, comparison of precision, recall and F1-score demonstrated that accuracy alone does not provide a complete description of model behaviour.

The findings support the use of multi-metric evaluation and explicit analysis of class imbalance in DR classification research. The study also establishes a methodological basis for the subsequent computational investigation planned in the corresponding thesis. Future work should apply a standardized experimental protocol to the APTOS 2019 dataset and evaluate selected transfer-learning architectures using the same preprocessing procedures, data partitioning strategy and performance measures.

References

1. Asiri, N., Hussain, M., Al Adel, F., & Alzaidi, N. (2019). Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey. *Artificial Intelligence in Medicine*, 99, 101701. <https://doi.org/10.1016/j.artmed.2019.07.009>
2. Bhulakshmi, D., & Rajput, D. S. (2024). A systematic review on diabetic retinopathy detection and classification based on deep learning techniques using

- fundus images. *PeerJ Computer Science*, 10, e1947.
3. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22), 2402–2410.
<https://doi.org/10.1001/jama.2016.17216>
 4. Khalifa, N. E. M., Loey, M., Taha, M. H. N., & Mohamed, H. N. E. T. (2019). Deep transfer learning models for medical diabetic retinopathy detection. *Acta Informatica Medica*, 27(5), 327–332.
<https://doi.org/10.5455/aim.2019.27.327-332>
 5. Raman, R., Srinivasan, S., Virmani, S., Sivaprasad, S., Rao, C., & Rajalakshmi, R. (2019). Fundus photograph-based deep learning algorithms in detecting diabetic retinopathy. *Eye*, 33(1), 97–109.
<https://doi.org/10.1038/s41433-018-0269-y>
 6. Tsiknakis, N., Theodoropoulos, D., Manikis, G., Ktistakis, E., Boutsora, O., Berto, A., Scarpa, F., Scarpa, A., Fotiadis, D. I., & Marias, K. (2021). Deep learning for diabetic retinopathy detection and classification based on fundus images: A review. *Computers in Biology and Medicine*, 135, 104599.
<https://doi.org/10.1016/j.combiomed.2021.104599>
 7. Voets, M., Møllersen, K., & Bongo, L. A. (2019). Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PLOS ONE*, 14(6), e0217541.
<https://doi.org/10.1371/journal.pone.0217541>