Canadian Journal of Marketing Research
ISSN: 0829-4836
https://canadian-jmr.com/index.php/cjmr

Research Article

Performance Evaluation of CNN and Hybrid Deep Learning Models for Retinal Disease Classification: A Review-Based Comparative Study

Dheeraj Kumar Mishra[1*] and Nidhi Sethi[2]
[1]Research Scholar, Renaissance University, Indore, India
[2]Associate Professor, Renaissance University, Indore, India

***Corresponding Author**
Dheeraj Kumar Mishra

Article History
Received: 18.08.2026
Accepted: 12.10.2026
Published: 22.11.2026

DOI: 10.61336/cjmr.1604.49

Abstract: Retinal diseases such as diabetic retinopathy (DR), age-related macular degeneration (AMD), glaucoma, and retinal vein occlusion (RVO) are leading causes of vision loss. Many of them cause no symptoms in the early stage, so regular screening is important. In many regions, however, there are too few eye specialists to screen everyone. This review compares two groups of deep learning models used to classify retinal images: standalone convolutional neural networks (CNNs) and hybrid models. Hybrid models add parts such as attention modules, recurrent layers, ensembles, transformers, or classical classifiers to a CNN. We followed the PRISMA 2020 reporting guideline to search, screen, and select studies. We compared the models on feature extraction, classification performance, robustness, computational cost, interpretability, and fitness for clinical use. Across the studies we reviewed, hybrid models usually report higher accuracy, better detection of small lesions, and more stable results on new data than standalone CNNs. The gain comes with extra computing cost, and many studies use different datasets and test rules, so their numbers cannot be compared directly. A well-tuned CNN backbone with a light attention module and transfer learning gives the best balance of accuracy, cost, and explainability for most screening settings. Heavier ensembles and CNN-transformer models suit cases where accuracy matters more than speed. Fair comparison needs patient-level data splits, external test sets, and shared reporting standards.

Keywords: Retinal Disease Classification, Deep Learning, Convolutional Neural Network, Resnet50, Hybrid Models, Medical Imaging, Artificial Intelligence

Copyright @ 2026: This is an open-access article distributed under the terms of the Creative Commons Attribution license which permits unrestricted use, distribution, and reproduction in any medium for non-commercial use (NonCommercial, or CC-BY-NC) provided the original author and source are credited.

INTRODUCTION

Losing sight changes a person's work, safety, and independence. Global estimates show that hundreds of millions of people live with some form of vision impairment [1]. Retinal diseases are a major part of this burden. About 103 million adults were estimated to have diabetic retinopathy in 2020, and the number is projected to reach about 160 million by 2045. Projections for AMD and glaucoma also point to a steep rise by 2040.

The main problem is that these diseases often stay silent until damage cannot be undone. Screening can find them early, but it needs trained graders and specialists. Countries with large diabetic populations and few eye doctors per person, India among them, find it hard to meet this need. Automated analysis of retinal images can help by sorting patients quickly, so that specialists see the people who need them most.

The retina can be photographed without any incision. Colour fundus photographs and optical coherence tomography (OCT) scans are the two most common image types. Deep learning, and CNNs in particular, can learn useful visual patterns directly from such images. A well-known early result came from Gulshan *et al*. [2], whose network detected referable diabetic retinopathy with very high accuracy on two independent image sets.

A single CNN still has weak points. It can miss tiny lesions, it treats every part of the picture alike, it struggles when one class is rare, and it often does worse on images from a new camera or hospital. To fix these problems, researchers

have built hybrid models that join a CNN with other components. There are now many such designs, but no clear guide that explains how they differ from plain CNNs and when the extra cost is worth it.

Research Questions

This review answers four questions.

- **RQ1:** Which CNN architectures are most often used for retinal disease classification, and how do they perform
- **RQ2:** What kinds of hybrid deep learning models have been proposed, and what problem does each one solve
- **RQ3:** How do hybrid models compare with standalone CNNs in accuracy, robustness, computing cost, and interpretability
- **RQ4:** What stops these models from being used safely in real clinics, and what should future research do about it

Contributions

- A clear grouping (taxonomy) of CNN and hybrid models for retinal image analysis
- A comparison across six dimensions instead of accuracy alone
- A checklist for judging study quality, which shows why published scores are often not directly comparable
- A set of practical recommendations for researchers and for hospitals that plan to adopt these tools

Background

Retinal Diseases

- **Diabetic Retinopathy:** High blood sugar damages the tiny blood vessels of the retina. Doctors look for microaneurysms (small bulges in vessels), bleeding spots, fatty deposits called hard exudates, and new abnormal vessels. A widely used scale sorts the disease into five levels, from no DR to proliferative DR.
- **Age-Related Macular Degeneration:** AMD harms the macula, the central part of the retina used for reading and recognising faces. The early sign is drusen, which are yellow deposits under the retina. The disease may progress to a dry form with tissue loss or a wet form with leaking new vessels
- **Glaucoma:** Glaucoma slowly damages the optic nerve. On a fundus photograph, doctors check the size of the optic cup relative to the optic disc and the thickness of the nerve rim
- **Retinal Vein Occlusion:** A blocked retinal vein causes bleeding, swollen veins, and fluid in the macula

Imaging Methods

From Hand-Made Features to Deep Learning: Earlier systems relied on features designed by engineers, such as vessel maps or texture measures, and then used a classifier such as a support vector machine. These pipelines took much effort and were sensitive to poor image quality. A CNN learns its own features. The first layers pick up edges and textures, and the deeper layers combine them into shapes such as lesions and the optic disc. Reviews of medical image analysis show that this ability is why CNNs became the standard tool (Table 1).

Review Methodology

We designed the review following the PRISMA 2020 guideline. This section lists the exact steps so that other researchers can repeat them.

Search Strategy

- **Databases:** Scopus, Web of Science, IEEE Xplore, PubMed, ScienceDirect, and SpringerLink
- **Period:** 2015 to 2026, which covers the deep learning era in retinal imaging
- **Search String (Adapted to Each Database):** ("retina" OR "retinal" OR "fundus" OR "OCT" OR "diabetic retinopathy" OR "glaucoma" OR "macular degeneration") AND ("deep learning" OR "convolutional neural network" OR "CNN" OR "hybrid" OR "attention" OR "ensemble" OR "transformer") AND ("classification" OR "grading" OR "detection")

Table 1: Retinal Imaging Methods

Method	What it shows	Best used for	Main limit
Colour fundus photography	Surface of the retina	Large-scale screening for DR and glaucoma	Cannot show layers beneath the surface
OCT	Cross-section of retinal layers	AMD, macular oedema	Costlier equipment
Fluorescein angiography	Blood flow and leakage	Vascular disease	Needs dye injection; rarely in public datasets

Table 2: Eligibility Criteria

Aspect	Include	Exclude
Publication	Peer-reviewed journal or conference paper, written in English	Editorials, abstracts only, theses
Task	Classification or grading of retinal disease from fundus or OCT images	Segmentation-only or image-quality-only studies
Method	CNN, or a CNN combined with another component	Methods with no deep learning
Data	Public dataset, or a clearly described private dataset	No dataset description
Results	Reports at least accuracy, or sensitivity and specificity, or AUC	No quantitative results

Table 3: Study Quality Checklist

Check	Question asked
Data split	Were images split by patient, not by image
External test	Was the model tested on data from a different source
Metrics	Were sensitivity, specificity, and AUC reported, not accuracy alone
Baseline	Was a fairly tuned baseline model included
Reproducibility	Are code, settings, and data sources shared

Studies that meet more checks carry more weight in our conclusions

Table 4: Public Retinal Image Datasets

Dataset	Image type	Approx. size	Labels	Source
EyePACS (Kaggle DR 2015)	Fundus	~88,000 images	5 DR grades	Cuadros and Bresnick
APTOS 2019	Fundus	~3,660 labelled images	5 DR grades	Kaggle competition
Messidor / Messidor-2	Fundus	1,200 / 1,748 images	DR grade, macular oedema risk	Decencière *et al.*
IDRiD	Fundus	516 images	DR grade, lesion maps, optic disc	Porwal *et al.*
ODIR-5K	Fundus	5,000 patients	8 disease groups	ODIR challenge
REFUGE	Fundus	1,200 images	Glaucoma, disc and cup masks	Orlando *et al.*
Kermany OCT	OCT	~84,000 images	CNV, DME, drusen, normal	Kermany *et al.*

Table 5: Main CNN Architectures

Model	Key idea	Parameters (approx.)	Typical role in retinal studies
VGG16	Stacks of small 3×3 filters	138 M	Early baseline; heavy to run
ResNet50	Shortcut connections that ease training of deep nets	25.6 M	Most common backbone and baseline
InceptionV3	Filters of several sizes in parallel	23.8 M	Used in early clinical-grade DR work
DenseNet121	Each layer reuses all earlier features	8 M	Good accuracy with fewer parameters
EfficientNet	Scales depth, width, and image size together	5.3 M (B0) and up	Strong on high-resolution fundus images
MobileNetV2	Light filters built for phones	3.5 M	Edge and smartphone screening

Eligibility Criteria

Data Extraction and Quality Checks: From each study we record the disease, dataset, image type, model design, training method, metrics, and whether external testing was done. We then score each study with a short quality checklist (Table 2,3).

Datasets and Evaluation Metrics

Public Datasets: Most studies train and test on a small set of public datasets. Knowing their size and labels helps explain why results differ from study to study (Table 4).

Three issues matter when reading results on these datasets. First, grade labels come from human graders who sometimes disagree, so the labels contain noise. Second, severe cases are far rarer than healthy ones, so the classes are imbalanced. Third, most datasets come from a few cameras and populations, which limits how well a model trained on them works elsewhere.

Metrics

Let TP, FP, TN, and FN be true positives, false positives, true negatives, and false negatives.

- **Accuracy:** is the share of all images classified correctly: (TP + TN) / (TP + TN + FP + FN)
- **Sensitivity (recall):** is the share of sick eyes that the model finds: TP / (TP + FN)
- **Specificity:** is the share of healthy eyes that the model clears: TN / (TN + FP)
- **Precision** is the share of positive calls that are correct: TP / (TP + FP)
- **F1-Score:** Is the balance of precision and recall
- **AUC:** is the area under the ROC curve. It measures how well the model separates sick from healthy eyes over all thresholds
- **Quadratic weighted kappa (QWK):** is the usual score for DR grading. It punishes a big grade error (for example, grade 0 given as 4) more than a small one [3]

$$\kappa_w = 1 - \frac{\sum_{i,j} w_{ij}\ O_{ij}}{\sum_{i,j} w_{ij}\ E_{ij}}, \qquad w_{ij} = \frac{(i-j)^2}{(N-1)^2}$$

Here O is the observed agreement matrix, E is the matrix expected by chance, and N is the number of grades.

In screening, a missed patient costs more than an extra referral, so sensitivity deserves more attention than accuracy. Accuracy alone can also look good when one class dominates the data.

CNN-Based Models (RQ1)

How a CNN Works: A CNN slides small filters over an image to detect patterns. Pooling layers shrink the maps, and a final layer turns the learned features into a class label. Because the same filter is used everywhere, a CNN can spot a haemorrhage wherever it appears in the picture. This built-in habit suits retinal lesions, which are small and can lie anywhere (Table 5).

ResNet50 deserves special mention. Its shortcut connections let the network grow deep without losing the training signal, and its size is moderate. For these reasons it is the default starting point in many retinal papers and the usual backbone inside hybrid designs.

Image size matters for retinal work. Microaneurysms are only a few pixels wide, so shrinking a fundus photograph too much can erase them. Models that accept larger inputs, such as the EfficientNet family, often benefit.

Main Architectures

Landmark Studies: The studies below showed that CNNs can reach screening-grade performance. The values are those reported by the original authors for their own data and test rules (Table 6).

These results are strong, but each one comes from a different dataset and test method. They show what is possible. They do not rank the models against each other (Table 7).

Strengths and Weaknesses of Standalone CNNs

Hybrid Models (RQ2): A hybrid model keeps a CNN as the eye of the system and adds another part to cover a known weakness. Table 8 summarises the main types, and the text below explains each one.

CNN with Attention

Attention modules let the network learn which channels or locations matter most. Squeeze-and-Excitation blocks re-weight feature channels. CBAM does this for channels and for spatial positions. In retinal images the highlighted areas often fall on the optic disc, macula, haemorrhages, and exudates, so attention can both raise accuracy on subtle lesions and give a visual hint of what the model used. The extra computing is small.

CNN with Recurrent Layers

In this design the CNN reads each image or patch, and an LSTM or GRU reads the resulting sequence. The sequence may be rows of a feature map, slices of an OCT volume, or a patient's visits over time. The idea is useful when order is part of the data. For a single flat fundus photograph, the gain over a good attention-based CNN is often small.

Table 6: Landmark CNN Studies (Values as Reported by the Original Authors)

Study	Disease and image	Reported result
Gulshan *et al.* [2]	Referable DR, fundus	AUC 0.991 and 0.990 on two test sets
Ting *et al.*	DR, glaucoma suspect, AMD, fundus	AUC about 0.94 for DR and glaucoma suspect, 0.93 for AMD
Abràmoff *et al.* [4]	DR in primary care, fundus	Sensitivity 87.2%, specificity 90.7% in a pivotal trial
De Fauw *et al.* [5]	Many retinal diseases, OCT	Referral decisions matched or beat retina specialists
Kermany *et al.*	CNV, DME, drusen, OCT	Accuracy above 96% using transfer learning
Burlina *et al.* [6]	AMD severity, fundus	Accuracy of about 88% to 92%
Li *et al.*	Glaucoma-like optic nerve damage, fundus	AUC 0.986

Table 7: Strengths and Weaknesses of Standalone CNNs

Strengths	Weaknesses
Learn features automatically	Can miss very small, faint lesions
Pre-trained weights are easy to reuse	Give equal weight to all image regions
Simple to build and deploy	Affected by class imbalance and by changes in camera or hospital
Moderate computing cost	See only local context in each filter
Mature software support	Hard to explain without extra tools

Table 8: Types of Hybrid Models and the Problem Each Addresses

Type	Added part	Problem addressed	Main cost
CNN + attention	SE, CBAM, or self-attention block	Focus on small or important regions	Small
CNN + recurrent	LSTM or GRU	Order in image patches, B-scans, or visits	Moderate
CNN + classical ML	SVM, random forest, boosting	Small datasets; overfitting	Low, but not end-to-end
Ensemble	Several CNNs combined	Individual model errors	High at inference
CNN + transformer	Self-attention over the whole image	Global context	High, needs more data

Table 9: Qualitative Comparison of Model Groups

Model group	Small-lesion detection	Class imbalance	Robustness on new data	Training and run cost	Explainability
Standalone CNN (VGG, ResNet, Inception, DenseNet)	Moderate	Low to moderate	Moderate	Low to moderate	Moderate (post-hoc tools)
Lightweight CNN (MobileNet)	Low to moderate	Low to moderate	Moderate	Low	Moderate
EfficientNet family	Moderate to high	Moderate	Moderate to high	Moderate	Moderate
CNN + attention	High	Moderate to high	Moderate to high	Low to moderate	High (attention maps)
CNN + recurrent	Moderate	Moderate	Moderate	Moderate to high	Moderate
CNN + classical ML	Moderate	Moderate	Moderate	Low	Moderate
Ensemble	High	High	High	High	Low
CNN + transformer	High	Moderate to high	High	High	Moderate

CNN with Classical Classifiers

Here a pre-trained CNN only extracts features, and a classifier such as an SVM or random forest makes the final decision. This can work well when labelled images are few, because training a full deep network on a small set tends to overfit. The downside is that the pipeline is no longer trained as one piece.

Ensembles

An ensemble averages or votes across several models, for example ResNet50, DenseNet121, and EfficientNet. Because the models rarely make the same mistake, the group is usually more accurate and more stable than any one member [7]. The price is longer run time and more memory, which can be a problem on low-cost devices.

CNN with Transformers

Vision transformers use self-attention to relate every part of an image to every other part [8]. They learn global patterns well but need a lot of training data. Hybrids use a CNN to find local lesions and transformer layers to judge how those lesions are spread across the retina, which matters for DR grading. Window-based variants such as the Swin Transformer reduce the cost.

Transfer Learning and Training Tricks

Transfer learning means starting from a network already trained on a large image set such as ImageNet, then fine-tuning it on retinal images. It helps because labelled medical data are scarce. Some work finds that the main benefit for medical images comes from good weight scaling rather than from the features themselves. Newer foundation models trained on huge sets of unlabelled retinal images show that this approach can also transfer well across tasks.

Other common tools are:

- **Preprocessing:** green-channel extraction, contrast enhancement, colour normalisation, and cropping of the black border.
- **Augmentation:** flips, rotations, and brightness changes that enlarge the training set
- **Handling imbalance:** class weights, focal loss, and oversampling methods such as SMOTE [9].
- **Ordinal learning:** treating DR grades as ordered numbers and training toward a QWK-friendly target

Comparative Analysis (RQ3)

Overall Comparison: The ratings below summarise the trends we see across the reviewed literature. They are qualitative on purpose, because the underlying studies use different data and test rules. Authors should support each rating with the studies in their own included-study table (Table 9).

Binary screening tasks, such as referable versus non-referable DR, give higher scores than five-grade grading. In grading, the middle grades (mild and moderate DR) are hard to tell apart even for human experts.

Table 10: Computing Cost of Model Groups

Aspect	Standalone CNN	Attention hybrid	Recurrent hybrid	Ensemble	CNN + transformer
Training cost	Low to moderate	Moderate	Moderate to high	High	High
Time per image	Low	Low to moderate	Moderate	High	Moderate to high
Memory	Low to moderate	Moderate	Moderate	High	High
Fit for phone or edge device	Good (light models)	Fair	Fair to poor	Poor	Poor

Accuracy and Sensitivity

Hybrid models, especially those with attention and those built as ensembles, usually report the best accuracy and the best recall for subtle lesions. The reason follows directly from the design. Attention puts the model's focus on small findings, and an ensemble averages out the mistakes of single networks.

Robustness and Generalisation

Single CNNs lose performance when the camera, lighting, or patient group changes. Ensembles and CNN-transformer hybrids tend to hold up better, but good evidence needs testing on data from a different hospital. Few studies do this. Gulshan *et al*. [9] is a useful example because it checked an algorithm against manual grading in two Indian eye hospitals.

Computing Cost

Explainability: A standalone CNN can be explained after the fact with heat-map methods such as Grad-CAM (or class activation maps. Attention-based hybrids show their focus directly, which is easier to present to doctors. Ensembles are hardest to explain because the explanations of several models must be merged. Heat maps can look convincing without being faithful to the real reasoning of the model, so they should be checked against lesions marked by clinicians.

Answer to RQ3

- Hybrid models usually give higher accuracy and better detection of small lesions than standalone CNNs
- Ensembles and CNN-transformer hybrids are the most robust, but they cost the most to run
- Standalone and light CNNs remain the practical choice for low-resource screening
- The best balance today is a strong backbone such as ResNet50 or EfficientNet, a light attention module, and transfer learning

DISCUSSION

Why Hybrid Models Tend to Do Better

Each hybrid part answers a specific weakness of a plain CNN. Attention fixes the habit of treating all regions alike. Ensembles reduce the random errors of single networks. Transformer layers add the wider view that local filters lack. Classical classifiers guard against overfitting when data are scarce. Because every improvement has a clear reason behind it, the higher scores reported for hybrids are believable and not just a matter of chance. Reviews of AI in retinal care reach a similar view of the field's direction.

How to Read Reported Results

The statement that hybrids beat CNNs should be read with care

- **Different Rules:** studies differ in dataset, image size, data split, and class definition, so their scores cannot simply be ranked
- **Data Leakage:** if both eyes of one patient, or several visits, fall into both the training and test sets, the score is inflated. A split by patient avoids this
- **Weak Baselines:** a new hybrid is sometimes compared with an under-tuned CNN, which exaggerates the gain
- **Publication Bias:** positive results are more likely to be published than failures
- **No outside Testing:** a high score on a held-out part of the same dataset does not show that the model works on other cameras or populations
- **These concerns are not unique to eye imaging.** They appear across medical imaging research that depends on limited labelled data [10].

Fitness for Clinical Use

Accuracy is Only One Part of a Safe Clinical Tool

- **Image quality:** images taken in busy primary-care clinics are not always good enough to grade, so an automatic quality check should come first.

Table 11: Main Challenges and Possible Responses

Challenge	Why it matters	Possible response
Limited, imbalanced data	Severe cases are rare, so models learn them poorly	Augmentation, focal loss, careful sampling
Label noise	Human graders disagree on middle grades	Several graders, adjudication, noise-robust training
Domain shift	Different cameras and populations change image appearance	Multi-centre training, domain adaptation
Explainability	Doctors need reasons, not only labels	Attention maps checked against clinician-marked lesions
Privacy	Sharing patient images across hospitals is restricted	Federated learning, anonymisation
Fairness	Uneven training data can cause uneven performance	Subgroup reporting, balanced data collection
Several diseases at once	Many models assume one label per image	Multi-label training
Weak clinical evidence	Most studies are retrospective	Prospective trials and real-world audits

- **Proof in real settings:** a pivotal trial in primary care showed that an autonomous DR system could work outside a research lab [4]. A study of automated grading software in a national screening programme also examined whether it was cost-effective against human graders
- **Local conditions:** AI applied to smartphone-based fundus photography has been tested for DR detection in India, which suggests that low-cost options can reach areas with few specialists
- **Trustworthy confidence:** modern networks are often over-confident. Calibrating the output probabilities helps doctors decide when to trust the result
- **Human oversight:** the model should act as a triage and support tool. Unclear cases should go to an ophthalmologist
- **Workflow:** the system should give a referral suggestion, a confidence value, and a visual explanation, and it should fit into existing hospital software

For rural screening, a light attention-based CNN on a portable camera may serve patients better than a heavier ensemble with slightly higher accuracy.

Ethics and Fairness

Retinal images are personal health data and need strong privacy protection. Models trained on one population may work less well on another, so reports should break results down by age, sex, and ethnicity wherever the data allow it.

Challenges and Limitations (RQ4)

Future Research Directions

- **Foundation models:** networks pre-trained on very large sets of unlabelled retinal images can be adapted to many tasks with few labels (Table 11)
- **Federated learning:** hospitals can train a shared model without exchanging patient images
- **Uncertainty estimates:** methods such as Monte Carlo dropout let a model signal when it is unsure [10]
- **Multimodal Models:** combining fundus images, OCT, and clinical data such as blood sugar and blood pressure may improve diagnosis and risk prediction
- **Validated Explanations:** explanation tools should be tested against lesions marked by clinicians, not only judged by eye
- **Light Hybrids:** compressed and quantised hybrids could bring attention-based accuracy to phones and portable cameras
- **Shared Benchmarks:** future datasets should offer patient-level splits, external test sets, and open evaluation code
- **Progression Models:** sequences of visits could be used to predict how a disease will change

CONCLUSION

This review compared standalone CNNs with hybrid deep learning models for retinal disease classification. CNNs such as ResNet50, InceptionV3, DenseNet, and EfficientNet are strong, well-understood baselines, and landmark studies show that they can reach screening-grade performance (RQ1). Hybrid models add attention, recurrent layers, classical classifiers, ensembles, or transformers, and each addition solves a specific weakness of plain CNNs (RQ2). In general, hybrids report higher accuracy, better detection of small lesions, and more stable results on new data, at the cost of more computing and more complexity (RQ3).

The evidence has limits. Studies use different datasets and test rules, some may leak data between training and test sets, and few have outside or prospective validation (RQ4). For practical use, a well-tuned CNN backbone with a light

attention module and transfer learning offers the best balance of accuracy, speed, and explainability. Ensembles and CNN-transformer hybrids fit settings where the best possible accuracy matters more than cost. Progress now depends on foundation models, federated and multimodal learning, validated explanations, and fair, standard evaluation.

REFERENCES

1. Bourne R.R.A. *et al.* "Trends in prevalence of blindness and distance and near vision impairment over 30 years: An analysis for the Global Burden of Disease Study." *The Lancet Global Health*, vol. 9, no. 2, 2021, pp. e130–e143.
2. Gulshan V. *et al.* "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." *JAMA*, vol. 316, no. 22, 2016, pp. 2402–2410.
3. Cohen J. "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit." *Psychological Bulletin*, vol. 70, no. 4, 1968, pp. 213–220.
4. Abràmoff M.D. *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, 2018, article 39.
5. De Fauw J. *et al.* "Clinically applicable deep learning for diagnosis and referral in retinal disease." *Nature Medicine*, vol. 24, 2018, pp. 1342–1350.
6. Burlina P.M. *et al.* "Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks." *JAMA Ophthalmology*, vol. 135, no. 11, 2017, pp. 1170–1176.
7. Ganaie M.A. *et al.* "Ensemble deep learning: A review." *Engineering Applications of Artificial Intelligence*, vol. 115, 2022, article 105151.
8. Dosovitskiy A. *et al.* "An image is worth 16x16 words: Transformers for image recognition at scale." *International Conference on Learning Representations (ICLR)*, 2021.
9. Chawla N.V. *et al.* "SMOTE: Synthetic minority over-sampling technique." *Journal of Artificial Intelligence Research*, vol. 16, 2002, pp. 321–357.
10. Cheplygina V., de Bruijne M., and Pluim J.P.W. "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis." *Medical Image Analysis*, vol. 54, 2019, pp. 280–296.
11. Gulshan V. *et al.* "Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India." *JAMA Ophthalmology*, vol. 137, no. 9, 2019, pp. 987–993.
12. Gal Y. and Ghahramani Z. "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning." *Proceedings of ICML*, PMLR 48, 2016, pp. 1050–1059.
13. Cuadros J. and Bresnick G. "EyePACS: An adaptable telemedicine system for diabetic retinopathy screening." *Journal of Diabetes Science and Technology*, vol. 3, no. 3, 2009, pp. 509–516.
14. Decencière E. *et al.* "Feedback on a publicly distributed image database: The Messidor database." *Image Analysis & Stereology*, vol. 33, no. 3, 2014, pp. 231–234.