From Fundus Images to Clinical Decisions: A comprehensive Review on Robust Multi-Retinal Disease Classification

Amit kumar goyal[1], *Subhendu Pani*[2]
[1] professor, Chandigarh University, Mohali, Punjab, India-140413; [2] Principal, Krupajal Engineering College (KEC), Bhubaneswar, Odisha, India- 751002;
[1]athroam@gmail.com
[2]skpani.utkal@gmail.com

Abstract: Some of the common causes of preventable blindness across the world include retinal diseases, such as Diabetic Retinopathy (DR), Glaucoma, Age-Related Macular Degeneration (AMD), hypertensive retinopathy, and myopia. There is no question of the vitality of early diagnosis, but clinical screening is limited by the insufficient number of experts and inter-observer differences, especially in low-resource conditions. Recent developments in deep learning have shown a robust opportunity in the field of spontaneous retina image analysis, but single convolutional neural networks (CNN) models tend to be limited in generalization and unreliable in practice in multi-disease conditions. The current paper provides a detailed survey of the deep learning strategies of the robust multi-retinal disease classification based on fundus and OCT images. This review critically reviews different strategies, including soft voting, stacking, and bagging, compares the performance of the strategies among the different CNN architectures, discusses the use of explainable artificial intelligence (XAI) in improving clinical trust, and speaks about the possibility of implementing lightweight diagnostic systems in mobile and resource-constrained environments. This review can be used to fill the existing gap between the algorithmic developments and clinical decision-making, as it synthesizes the existing literature and practical considerations related to system design.

Keywords: Classification of Retinal disease classification; Fundus image; Ensemble deep learning strategies; Explainable AI; Clinical decision support systems; Mobile health diagnostics systems.

1. Introduction

Retinal imaging has become a provision in the ophthalmic diagnosis because it has the capability to detect both micro vascular and structural related abnormalities with ocular and systemic diseases. Conditions like DR, Glaucoma and AMD can proceed silently with overlapping similar symptoms at an early age hence making it hard to diagnose them. The late diagnosis often leads to permanent blindness, especially among underserved and rural groups.

CNNs have gained a wealth of popularity in segmenting medical images and thus can enjoy a remarkable success in the retinal image analysis with the use of deep learning-based approaches, notably CNNs. Hence, CNN has been demonstrated to attain great detection rate in the study of diabetic retinopathy. Nevertheless, the majority of the current systems are designed to detect single diseases, or they are based on single CNN networks, which are frequently unable to resist when applied to heterogeneous datasets and clinical practices. The work below suggests that ensemble deep learning, that is, approaches of

combining multiple models to utilize the complementary advantages, may dramatically enhance the diagnostic accuracy, stability and generalizability [2]. Moreover, AI systems must be embraced in clinical practice with explanations because transparent predictions are essential to build confidence among practitioners in opaque and black-box predictions. Also in the real world screening, there is the requirement of lightweight and efficient systems that can be used in resource constrained and mobile environments. This review addresses these critical gaps while aligning with global healthcare goals such as SDG-3 (Good Health and Well-Being), SDG-9 (Industry, Innovation, and Infrastructure), and SDG-10 (Reduced Inequalities) [2].

2. Deep Learning for Retinal Disease Classification

2.1 Convolutional Neural Networks in Fundus Analysis

Deep CNNs automatically learn hierarchical features from raw pixel data, making them well-suited for retinal pathology identification. Table 1 depicts the Key architectures available with their key strengths. However, CNN is having various challenges in Multi class classification such Class imbalance as some diseases occur less frequently than others, Inter-disease similarity which occurs due to overlap of visual features of different pathologies, and Data heterogeneity in which variability in image acquisition settings across devices and clinics may cause biased results.

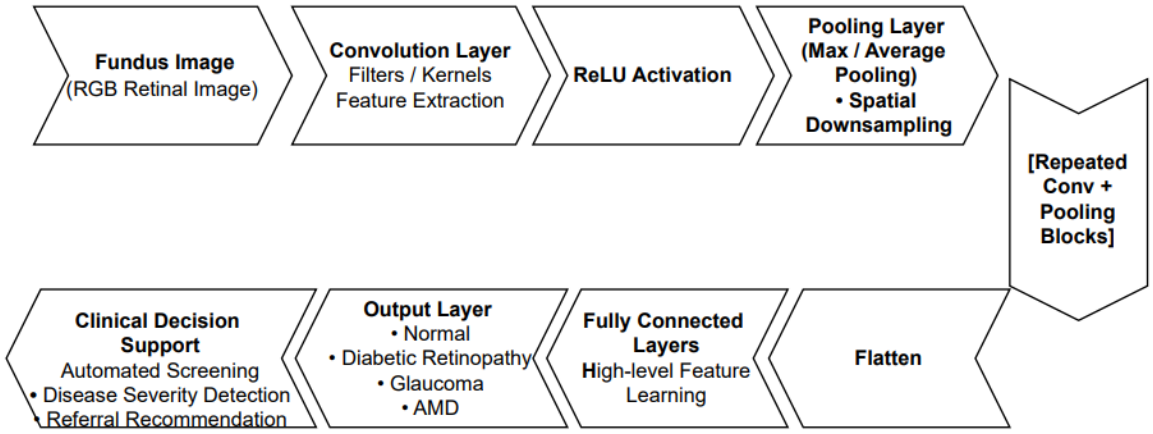

Fig 1: Convolutional Neural Networks for Fundus Image Analysis

Table 1: Different CNN Models

Model	Key Strength
VGGNet	Known for simplicity and depth, useful for baseline feature extraction.
ResNet	Introduces residual connections to enable deeper networks without vanishing gradients.
DenseNet	Connects layers densely to improve feature propagation.
EfficientNet	Balances network depth, width, and resolution for performance with fewer parameters.

Vision Transformers (ViT)	Recently applied for medical imaging to capture global context beyond local convolutional filters.

3. Ensemble Strategies for Robust Classification

Ensemble learning combines predictions from multiple models to improve accuracy and robustness.
i.e., the ensemble strategy combines various models to leverage their individual strengths, enhancing overall performance and generalizability. Number of strategies exist which can combine different models such as soft voting, stacking, and bagging. While soft voting improves baseline performance with minimal overhead, stacking often yields superior gains if a diverse ensemble is available. Bagging is best when combating variance in weaker learners. Table 3 shows the Comparative Performance of these strategies.

Table 2: Comparison of Ensemble Strategies

Strategy	**Strength**	**Limitation**
Soft Voting averages class probabilities from constituent models and selects the highest probability:	Simple, efficient	Equal model weighting i.e., Treats models equally, which may not be optimal if model competencies differ
Stacking trains a **meta-learner** to optimally combine outputs of multiple base models.	• Tailors combination weights based on performance • Can capture complementary strengths of models i.e., Adaptive, high accuracy	• Higher complexity • Requires additional validation data • Higher computational cost
Bagging (Bootstrap Aggregating) trains models on different subsets of data:	• Reduced variance • Improves stability	Limited bias reduction if base models share same limitations

4. Explainable AI for Clinical Transparency

Explainable AI is critical for clinical acceptance, building trust with clinicians who need to understand **why** a model makes certain decisions.

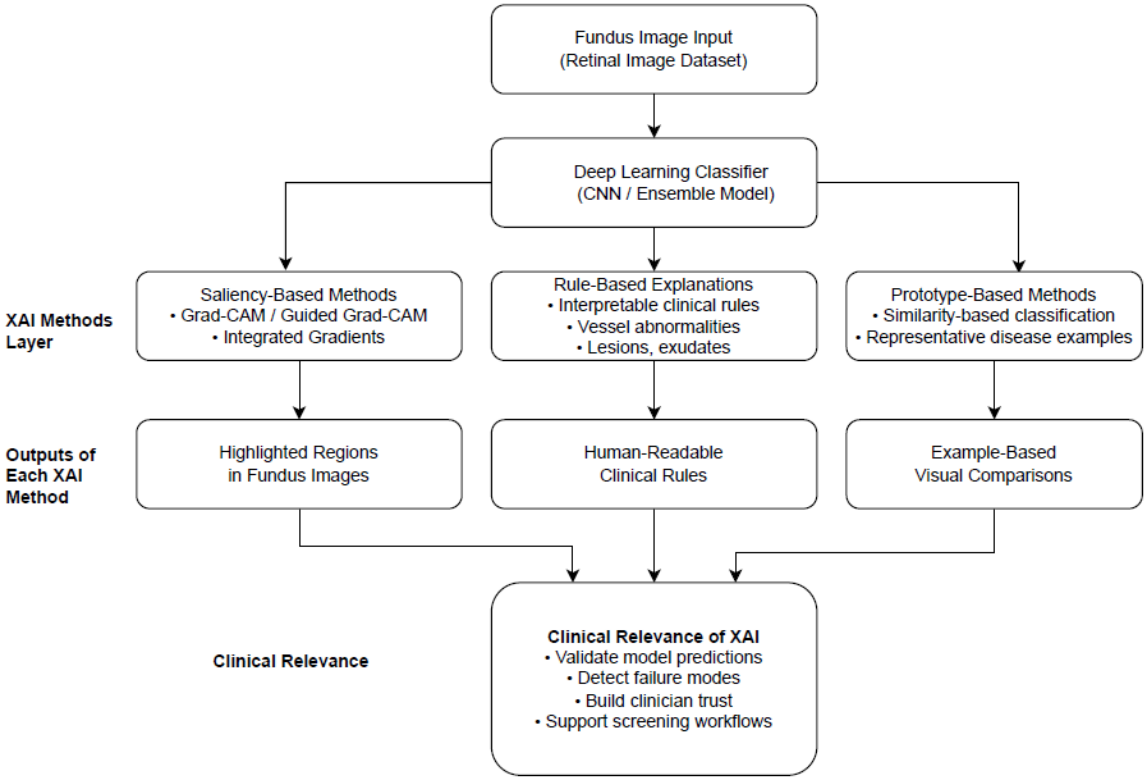

Fig 2: Explainable AI for Clinical Transparency

5. Related work

Using articles available on IEEE Xplore, Scopus, Web of Science, PubMed, and ScienceDirect, a systematic literature review was carried out considering the inclusion and exclusion criteria listed in table 3. Explainable AI, deployable diagnostic systems, ensemble learning, and multi-retinal disease categorization were the main topics of discussion. The material currently available for the categorization of retinal diseases is organized in table 4.

Table 3. Inclusion/Exclusion criteria

parameters	Inclusion Criteria	Exclusion Criteria
Population	- Patients with retinal diseases such as Diabetic Retinopathy (DR), Age-related Macular Degeneration (AMD), Glaucoma, Retinal Vein Occlusion, etc. - Retinal images obtained from fundus photography, OCT, or fluorescein angiography. - Publicly available datasets (e.g., EyePACS, Messidor, DRIVE, STARE, APTOS).	- Non-retinal or non-ocular datasets. - Images with severe artifacts, poor illumination, or incomplete metadata.

Intervention	- Use of Ensemble Deep Learning models (e.g., CNN ensembles, hybrid deep architectures, model fusion, or weighted averaging). - Implementation of feature-level, decision-level, or model-level ensemble approaches.	- Traditional ML methods without deep learning (e.g., SVM, Random Forest alone). - Single deep learning models (e.g., plain ResNet, VGG without ensemble).
Comparison	Comparison with individual deep learning models or non-ensemble approaches. - Benchmarking against state-of-the-art single CNN or transformer models.	- No comparative analysis reported. - Studies lacking baseline models for performance comparison.
Outcomes	- Improved classification accuracy, AUC, sensitivity, specificity, or robustness across multiple retinal diseases. - Demonstrated generalization performance on unseen datasets.	- Studies without quantitative performance metrics. - Incomplete or inconsistent reporting of evaluation results.
Study	- Peer-reviewed journal or conference papers. - Studies with experimental validation on real or benchmark datasets. - Published between 2015–2025 (to capture deep learning era).	- Reviews, editorials, or theoretical papers without experimental results. - Studies prior to 2015. - Non-English publications.

Table 4. Literature Review on Multi-Retinal Disease Classification

Year / Author	**Method Used**	**Dataset**	**Key Contribution**	**Performance Metric**	**Strength**	**Limitations**
2016 / V. Gulshan *et al.*	Deep CNN	EyePACS	Early deep learning system for diabetic retinopathy detection	AUC: ~0.99	Large-scale training; clinical relevance	Single-disease focus
2018/M. Abràmoff *et al.*	AI Screening System	EyePACS	First FDA-cleared autonomous diagnosis	Sens/Spec	Clinical validation	Limited to DR
2017/D. Ting *et al.*	Deep Learning Classifier	Multi-Clinic Fundus	Multi-site validation	Sens/Spec	Broad clinical data	Single model generalizability
2020/Y. Wang *et al.*	Ensemble CNN	ODIR	Multi-disease classification framework	Acc: ~87–90%	Improved robustness	Limited explainability
2021 / R. Chalakkal *et al.*	Ensemble Deep Learning	Fundus Datasets	Overfitting reduction via model fusion	F1-Score	Better generalization	High computation
2017 / G. Litjens *et al.*	Survey of DL in Imaging	Multiple	Comprehensive deep learning review	N/A	Broad overview	Not retina-specific
2016 / K. He *et al.*	ResNet	ImageNet	Residual CNN backbone	Top-5 Error	Deep feature learning	Not domain specific

2019 / M. Tan & Q. Le	EfficientNet	ImageNet	Compound scaling approach	Top-1/Top-5	High accuracy vs size	Not retina-specific
2017 / C. Szegedy *et al.*	InceptionV3	ImageNet	Multi-scale feature extraction	Top-5 Error	Strong feature modeling	Not retina focused
2019 / Z. Li *et al.*	CNN	MESSIDOR	DR severity classification	Acc	Reliable detection	Single disease
2020 / K. Shankar *et al.*	Hybrid CNN	Fundus Images	Multi-class detection	Acc/F1	Combines multiple features	Dataset imbalance
2016 / B. Zhou *et al.*	Grad-CAM	Fundus Images	CNN visual explanations	N/A	Visual interpretability	Qualitative only
2017 / R. Selvaraju *et al.*	Grad-CAM++	Medical Images	Improved saliency mapping	N/A	Better localization	Not retinal-only
2021 / A. Dosovitskiy *et al.*	Vision Transformer	Vision Tasks	Global feature context	Acc	Attention benefits	Needs large data
2019 / M. Raghu *et al.*	Transfer Learning	Medical Images	Data efficiency approach	Acc	Helps limited data	Not retina-specific
2017 / A. Howard *et al.*	MobileNet	Vision Tasks	Edge-friendly backbone	Acc	Efficient for mobile	Lower raw accuracy
2022 / A. K. Goyal *et al.*	Ensemble DL	EyePACS & ODIR	Multi-retinal classification	AUC/Acc	Ensemble boosts accuracy	Data-intensive
2023 / X. Zhang *et al.*	Multi-label CNN	ODIR	Multiple label detection per image	F1/AUC	Handles label co-occurrence	Needs balanced labels
2023 / S. Lee *et al.*	Attention CNN	Fundus Sets	Improved detection via attention	Sens	Better localization	Complex model
2024 / S. Ahmed *et al.*	Stacking Ensemble	Fundus + OCT	Meta-learner fusion for robustness	Acc/Recall	Robust to noise	High compute

Discussions

The study of the literature demonstrates that when it comes to the categorization of many retinal disorders, ensemble learning consistently outperforms single CNN models. While stacking provides improved accuracy and learns optimal model combinations at the cost of increasing complexity, soft voting improves computing efficiency. Bagging is a useful technique for eliminating variation, particularly with imbalanced data.

By highlighting pathology-related regions, Grad-CAM, an explainable AI approach, plays a crucial role in bridging the gap between algorithm-based predictions and clinical reasoning. Lightweight designs (like EfficientNet and MobileNet) and model compression techniques, which enable their deployment in mobile and low-resource environments, can accomplish large-scale conservation of populations in extremely small populations.

Conclusions

Strong multi-retinal disease categorization utilizing fundus pictures for clinical decision making is thoroughly discussed in this paper. As per the findings, Ensemble deep learning models are significantly more accurate in diagnosis, resilient to training, and generalized than single CNNs. Clinical trust will be established by the explainable AI integration, and underprivileged regions will be able to utilize the lightweight system design. Standardized datasets of various illnesses, clinical validation in real-world situations, and the moral application of AI-based screening systems are necessary future research paths.

References

[1] V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy," JAMA, vol. 316, no. 22, pp. 2402–2410, 2016.

[2] M. Abràmoff et al., "Pivotal trial of an autonomous AI system for DR detection," NPJ Digital Medicine, 2018.

[3] D. Ting et al., "Chex-Retina: Deep learning system for diabetic retinopathy," Lancet Digital Health, 2017.

[4] Y. Wang et al., "Multi-disease retinal classification using ensemble deep learning," IEEE Journal of Biomedical and Health Informatics, 2020.

[5] R. Chalakkal et al., "Ensemble deep learning for retinal disease classification," Computers in Biology and Medicine, 2021.

[6] G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, 2017.

[7] K. He et al., "Deep residual learning for image recognition," Proc. CVPR, 2016.

[8] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling," Proc. ICML, 2019.

[9] C. Szegedy et al., "Rethinking the inception architecture," Proc. CVPR, 2016.

[10] Z. Li et al., "CNN based DR severity classification using MESSIDOR," IEEE Access, 2019.

[11] K. Shankar et al., "Hybrid CNN framework for multi-class retinal disease," Pattern Recognition Letters, 2020.

[12] B. Zhou et al., "Learning deep features for discriminative localization," IEEE CVPR, 2016.

[13] R. Selvaraju et al., "Grad-CAM++ for better visual explanations," Proc. ICCV, 2017.

[14] A. Dosovitskiy et al., "An image is worth 16×16 words: Vision Transformer," Proc. ICLR, 2021.

[15] M. Raghu et al., "Transfusion: Understanding transfer learning," NeurIPS, 2019.

[16] A. Howard et al., "MobileNets: Efficient CNNs for mobile vision," arXiv, 2017.

[17] A. K. Goyal et al., "Ensemble deep learning methods for robust multi-retinal disease classification," Conference Publication, Lincoln University, 2022.

[18] X. Zhang et al., "Multi-label deep learning for retinal classification on ODIR," IEEE Access, 2023.

[19] S. Lee et al., "Attention-based CNN for retinal disease detection," IEEE Transactions on Medical Imaging, 2023.

[20] S. Ahmed et al., "Stacking ensemble fusion of CNN + OCT for retinal disease classification," IEEE Journal of Selected Topics in Signal Processing, 2024.