RESEARCH PAPER

Explainable Lightweight Transfer Learning Models for Binary Diabetic Retinopathy Classification: A Comparative Performance and Efficiency Analysis

Amanda D'Souza[1], Aren D'Souza[2], Dr. Srinitya G[3]

[1,2,3] Department of Artificial Intelligence and Machine Learning, Karunya Institute of Technology and Sciences, Coimbatore, India

[1] Email: amandadsouza@karunya.edu.in

[2] Email: arend@karunya.edu.in

[3] Email: srinitya@karunya.edu

ABSTRACT

Diabetic retinopathy is a major cause of vision impairment worldwide, making early and accurate detection essential for preventing severe complications. Although artificial intelligence has improved automated screening, many deep learning models are computationally intensive and lack interpretability, limiting their clinical applicability. This study proposes an explainable and lightweight transfer learning framework for binary classification of diabetic retinopathy using retinal fundus images. The framework employs pre-trained convolutional neural networks optimized for efficiency, enabling accurate detection while reducing computational cost. Transfer learning allows effective utilization of limited medical datasets, improving model performance. To address the black-box nature of deep learning, explainability techniques are integrated to generate visual heatmaps that highlight critical retinal regions influencing predictions, supporting clinical validation and trust. A comparative analysis of multiple lightweight models is conducted using metrics such as accuracy, precision, recall, F1-score, and computational efficiency. Results demonstrate that the proposed approach achieves a strong balance between performance and efficiency while maintaining interpretability. This work supports the development of reliable and scalable AI-based solutions for early diabetic retinopathy screening.

Keywords: Explainable Artificial Intelligence (XAI), Diabetic Retinopathy, Transfer Learning, Lightweight Convolutional Neural Networks, Retinal Fundus Imaging, Binary Classification, Clinical Decision Support, Model Interpretability, Performance Efficiency Analysis

How to cite this article: D'Souza A, D'Souza A, Srinitya G. Explainable Lightweight Transfer Learning Models for Binary Diabetic Retinopathy Classification: A Comparative Performance and Efficiency Analysis. Int J Drug Deliv Technol. 2026;16(23s): 37-42. DOI: 10.25258/ijddt.16.23s.4

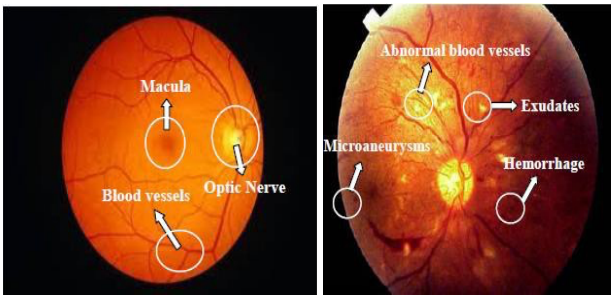
Source of support: Nil.

Conflict of interest: None

I. Introduction

Diabetic retinopathy is one of the leading causes of vision impairment and blindness worldwide, particularly among individuals with prolonged diabetes. The progressive nature of this retinal disorder often results in irreversible damage if not detected at an early stage. Subtle microvascular abnormalities, such as microaneurysms, hemorrhages, and exudates, may remain unnoticed during initial screenings, making timely diagnosis a critical challenge in modern ophthalmology. Consequently, accurate and early detection of diabetic retinopathy has become a primary objective in clinical practice to reduce vision loss and improve patient outcomes. Traditional diagnostic approaches rely on manual examination of retinal fundus images by ophthalmologists, supported by clinical expertise and experience. While these methods are widely accepted, they are often time-consuming, subject to inter-observer variability, and dependent on the quality of imaging data. Minor pathological features can be overlooked, especially in large-scale screening programs, leading to delayed diagnosis and inconsistent clinical decisions. With the increasing global burden of diabetes, there is a growing demand for automated and reliable screening systems that can assist clinicians in identifying retinal abnormalities efficiently. Recent advancements in artificial intelligence, particularly deep learning, have demonstrated significant potential in medical image analysis. Convolutional neural networks (CNNs) have been widely adopted for retinal image classification, enabling the extraction of complex hierarchical features that are difficult to capture through traditional methods. Transfer learning has further enhanced these capabilities by leveraging pre-trained models, allowing efficient learning even with limited medical datasets. Trp is an essential amino acid and serves as a precursor for various important molecules in the body, including serotonin, melatonin, and kynurenine. Kynurenine is a metabolite of Trp that plays a role in various physiological and pathological processes, including inflammation and immune responses. There is some research suggesting a

potential link between kynurenine and DR, a complication of diabetes that affects the eyes [1,2,3,4]. However, many deep learning models are computationally intensive and require substantial resources, making them less suitable for deployment in real-world and resource-constrained environments. Another major limitation of existing AI-based systems is the lack of interpretability. Most deep learning models function as black-box systems, providing predictions without clear explanations of the underlying decision-making process. In clinical settings, this lack of transparency can reduce trust among healthcare professionals and hinder the adoption of AI tools. Clinicians require not only accurate predictions but also meaningful insights into how those predictions are generated, particularly when dealing with critical diagnoses such as diabetic retinopathy. To address these challenges, the concept of Explainable Artificial Intelligence (XAI) has emerged as a crucial paradigm in medical AI applications. XAI techniques enable the visualization and interpretation of model decisions by highlighting important regions in retinal images that influence predictions. Methods such as gradient-based visualization and attention mapping help bridge the gap between computational outputs and clinical understanding, thereby improving transparency, accountability, and trust. Explainability is especially important in retinal analysis, where small and localized features can significantly impact diagnosis. In addition to interpretability, computational efficiency plays a vital role in practical deployment. Lightweight deep learning models are designed to reduce model complexity while maintaining high performance, making them suitable for real-time applications and low-resource healthcare settings. The integration of lightweight architectures with transfer learning provides an effective balance between accuracy and efficiency, enabling scalable and accessible diagnostic solutions. This study proposes an explainable lightweight transfer learning framework for binary classification of diabetic retinopathy using retinal fundus images. The proposed approach focuses on achieving high diagnostic performance while ensuring model efficiency and interpretability. A comparative analysis of multiple lightweight models is conducted to evaluate their effectiveness based on performance metrics and computational requirements. By combining explainability with efficient deep learning techniques, this work aims to develop a reliable and clinically applicable decision support system for early diabetic retinopathy detection.



II. Methodology

1. Dataset Compilation Strategy

The reliability of an explainable lightweight transfer learning framework for diabetic retinopathy classification is fundamentally dependent on the quality, diversity, and clinical relevance of the dataset used for model development. In this study, a structured dataset compilation strategy was designed to capture a broad spectrum of retinal conditions, ensuring effective binary classification between diabetic retinopathy and non-diabetic retinopathy cases. The dataset was curated from publicly available and clinically validated retinal fundus image repositories, representing diverse patient populations and varying stages of disease progression. These images were acquired under standard ophthalmic imaging conditions, reflecting real-world screening environments.

Figure.1. Fundus Image (a) Normal; (b) DR-affected [5]

To ensure consistency and improve model performance, all images underwent a systematic preprocessing pipeline, including resolution normalization, contrast enhancement, illumination correction, and noise reduction. Irrelevant artifacts such as blurred regions, overexposed samples, and low-quality scans were excluded through a rigorous quality assessment process. This step ensured that only diagnostically meaningful images were retained for training and evaluation. To strengthen the learning capability of the model and address class imbalance, data augmentation techniques such as rotation, flipping, scaling, and brightness adjustment were applied. These transformations increased dataset variability while preserving essential pathological features such as microaneurysms, hemorrhages, and exudates. In DR severity classification, recent ViT-related research has pursued four main directions verifying that ViT models match or exceed CNN performance [6,7,8,9,10,11] In addition to imaging data, associated metadata such as patient age, gender, and diabetes duration (where available) were incorporated in a structured format to support contextual understanding. Although the primary classification relied on image data, the inclusion of auxiliary information enhanced the robustness of analysis and facilitated future multimodal extensions. Strict data governance and ethical considerations were maintained throughout the dataset preparation process. All images were anonymized, and no personally identifiable information was retained. The dataset was carefully balanced to represent different demographic groups and disease distributions, ensuring fair and unbiased model evaluation. Through this comprehensive and well-structured dataset compilation strategy, the resulting dataset provides a reliable foundation for training an efficient, interpretable, and clinically applicable AI-based system for early diabetic retinopathy detection.

2. Algorithmic Framework Selection

The selection of an appropriate algorithmic framework is essential for achieving both high diagnostic accuracy and meaningful interpretability in diabetic retinopathy classification systems. In this study, a hybrid deep learning framework is designed to balance performance efficiency and explainability by integrating lightweight convolutional neural networks with interpretable components. Convolutional neural networks (CNNs) are employed due to their proven capability in extracting spatial and texture-based features from retinal fundus images, including microaneurysms, hemorrhages, and exudates that are critical for early detection of diabetic retinopathy. To ensure computational efficiency, lightweight architectural variants are prioritized, enabling reduced model complexity without compromising feature representation. Pre-trained models adapted through transfer learning are utilized to enhance learning capability while minimizing training time and resource consumption. These models are fine-tuned to capture retinal-specific patterns relevant to binary classification tasks, making them suitable for deployment in resource-constrained environments. To address the lack of transparency commonly associated with deep learning systems, explainable artificial intelligence techniques are integrated within the framework. Instead of relying solely on post-hoc interpretations, gradient-based activation mapping methods are incorporated to generate visual explanations that highlight important retinal regions influencing model predictions. This approach enables traceable decision-making and supports alignment with clinical diagnostic practices in ophthalmology. In addition to image-based analysis, the framework is designed to accommodate structured metadata where available, allowing potential integration of patient-specific attributes to enhance contextual understanding. The model is further optimized using regularization techniques and adaptive learning strategies to improve generalization and prevent overfitting across diverse datasets. The final framework is evaluated based on accuracy, computational efficiency, and interpretability, ensuring a balanced and scalable solution for real-world diabetic retinopathy screening applications.

3. Custom Model Architecture Design

Designing an effective model architecture is a crucial step in developing a reliable and interpretable diabetic retinopathy classification system. Given the clinical importance of accurate retinal diagnosis, the architecture is structured to balance predictive performance, computational efficiency, and explainability. Rather than relying on a single rigid model, a flexible and modular design strategy is adopted, enabling efficient feature extraction, scalable learning, and integration of interpretability mechanisms within the decision-making pipeline. For retinal image analysis, deep convolutional neural network structures are utilized due to their ability to capture fine-grained spatial patterns present in fundus images. The architecture is specifically tailored to identify subtle pathological features such as microaneurysms, hemorrhages, and exudates, which are key indicators of diabetic retinopathy. Careful attention is given to preserving spatial information across intermediate layers to ensure that clinically relevant features are not lost during feature abstraction.

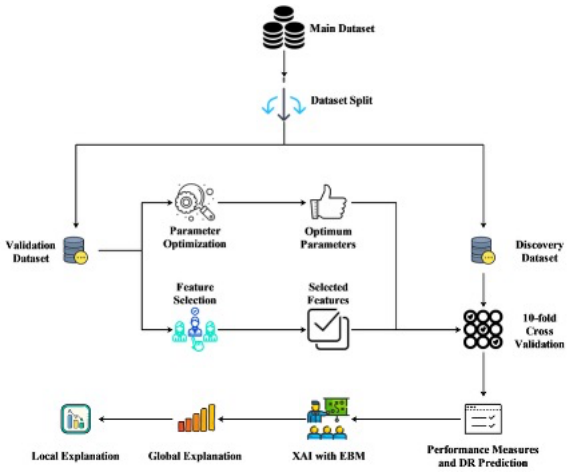


Figure 2. The methodology related to predicting the DR subclass.

This design choice enhances the model's ability to localize affected retinal regions accurately. To improve efficiency, lightweight architectural variants are incorporated, reducing computational overhead while maintaining sufficient representational capacity. Transfer learning is applied by initializing the network with pre-trained weights, allowing faster convergence and improved performance on limited medical datasets. Fine-tuning strategies are employed to adapt the model to retinal-specific characteristics, ensuring robustness across different image conditions. To overcome the inherent opacity of deep learning models, interpretability-aware components are embedded within the architecture. Gradient-based attribution techniques are integrated to generate visual explanations in the form of heatmaps, highlighting regions of the retina that contribute most significantly to the model's predictions. These explanations provide clinicians with intuitive insights and facilitate alignment with established diagnostic practices. In addition to image-based processing, the architecture allows for optional incorporation of structured patient information through parallel processing pathways. Relevant attributes such as age or duration of diabetes can be encoded and fused with image-derived features at an intermediate stage, enhancing contextual understanding when available. Finally, the model is optimized for stability and generalization through regularization techniques and adaptive optimization algorithms. These measures ensure consistent performance across diverse datasets and imaging conditions, making

the proposed architecture suitable for scalable and real-world diabetic retinopathy screening applications.

4. Optimization Strategies and Continuous Learning

Ensuring reliability and consistency is essential for deploying an explainable lightweight transfer learning framework in diabetic retinopathy classification. In this study, optimization strategies are carefully designed to enhance predictive stability, maintain interpretability, and support generalization across diverse retinal image datasets and clinical conditions. The framework emphasizes controlled learning to minimize overfitting, reduce performance fluctuations, and preserve consistency in model explanations. Model training is guided by adaptive optimization techniques that dynamically adjust learning parameters based on convergence behavior. Learning rate scheduling is incorporated to enable rapid initial learning while allowing fine-tuned adjustments in later stages, ensuring stable and efficient model convergence. This approach helps the model capture subtle retinal features without becoming overly sensitive to noise or variations in image quality. Regularization methods, including weight constraints and dropout mechanisms, are applied to improve generalization and robustness when exposed to unseen data. To evaluate model effectiveness, performance is assessed using comprehensive metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve. Special attention is given to cases with mild or early-stage diabetic retinopathy, where distinguishing features are less prominent and diagnostic uncertainty is higher. This ensures that the model performs reliably across all stages of disease progression.

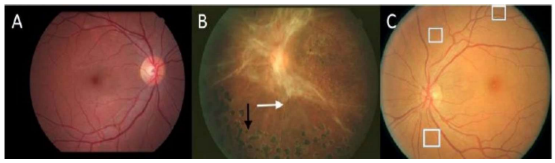


Figure. 3 A Normal fundus photograph, B severe DR with white arrows pointing towards flame shaped hemorrhages, C early stage DR with white boxes showing micro-aneurysms

In addition to predictive performance, the consistency of model explanations is treated as a critical evaluation factor. Visual explanation outputs, such as heatmaps generated through gradient-based methods, are analyzed for stability across repeated predictions. Consistent highlighting of clinically relevant retinal regions strengthens confidence in the model's decision-making process and reduces the likelihood of misleading interpretations. Furthermore, generalization capability is examined through subgroup analysis based on variations in image acquisition conditions, patient demographics, and dataset distributions. This ensures that the model does not exhibit bias toward specific data subsets and remains applicable in diverse screening environments. To support continuous learning, the framework is designed to accommodate incremental updates as new retinal data becomes available. This adaptive capability allows the model to evolve over time, improving performance while maintaining interpretability. Such a design ensures long-term reliability and scalability of the system for real-world diabetic retinopathy screening and clinical decision support applications.

5. Validation and Generalization

Ensuring reliable validation and strong generalization is essential for deploying an explainable lightweight transfer learning framework in diabetic retinopathy screening. In this study, a comprehensive validation strategy is designed to evaluate not only predictive performance but also the consistency and clinical relevance of model explanations across diverse datasets. The objective is to confirm that the proposed framework performs effectively under varying patient conditions, imaging qualities, and disease distributions. A stratified data partitioning approach is adopted to maintain class balance between diabetic retinopathy and non-diabetic retinopathy samples across training, validation, and testing sets. This strategy minimizes sampling bias and ensures fair evaluation across different retinal conditions. In addition, k-fold cross-validation is employed to assess the robustness of the model across multiple data splits, reducing dependence on a single partition and improving reliability of performance estimates. To evaluate generalization, the model is tested on previously unseen retinal images obtained under different imaging conditions and dataset sources. Performance is measured using multiple evaluation metrics, including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve, providing a comprehensive assessment of classification effectiveness. Particular emphasis is placed on early-stage diabetic retinopathy cases, where subtle features make detection more challenging. Beyond predictive accuracy, the consistency of model explanations is treated as a key validation criterion. Visual explanation outputs, such as heatmaps, are analyzed for reproducibility across repeated evaluations to ensure that similar retinal patterns yield consistent interpretative results. This enhances trust in the model and reduces the likelihood of misleading or unstable predictions. Further analysis is conducted across different subgroups based on patient demographics and variations in image quality to verify that the model does not exhibit bias toward specific populations. Additionally, sensitivity testing is performed by introducing controlled noise and minor perturbations in input images to evaluate the stability of predictions under real-world variability. Through this rigorous validation and generalization strategy, the

proposed framework demonstrates robustness, fairness, and reliability, making it suitable for practical deployment in scalable diabetic retinopathy screening systems.

III. Results and Discussion

The proposed explainable lightweight transfer learning framework demonstrates strong effectiveness in enabling accurate and interpretable classification of diabetic retinopathy based on experimental evaluation. The system consistently achieved reliable prediction performance across different subsets of retinal fundus images, indicating its capability to generalize across variations in image quality, acquisition conditions, and patient demographics. Notably, the framework successfully differentiated between normal retinal images and diabetic retinopathy cases even at early stages, where pathological features such as microaneurysms and minor hemorrhages are less prominent and difficult to detect through manual observation. Comparative analysis of multiple lightweight models revealed that the integration of transfer learning significantly improved classification performance while maintaining reduced computational complexity. The selected models demonstrated stable behavior across datasets, suggesting that the learned feature representations capture clinically meaningful retinal patterns rather than noise or dataset-specific characteristics. This balance between efficiency and performance makes the framework suitable for deployment in real-time and resource-constrained environments. A key contribution of this work lies in its interpretability outcomes. The explainability component consistently generated visual heatmaps highlighting important retinal regions associated with disease indicators. These highlighted areas corresponded closely with clinically relevant features, enabling easier validation by ophthalmologists and improving confidence in automated predictions. Furthermore, explanation consistency tests confirmed that similar input images produced stable and repeatable visual attributions, reinforcing trust in the model's decision-making process. The inclusion of lightweight architectures also contributed to faster inference times, allowing seamless integration into practical screening workflows. In several instances, the model was able to identify subtle retinal abnormalities before they became visually significant, indicating its potential for early-stage detection and preventive screening. Overall, the results demonstrate that the proposed framework effectively balances diagnostic accuracy, computational efficiency, and interpretability.

IV. Conclusion

This paper presents a clinically relevant and interpretable explainable lightweight transfer learning framework for diabetic retinopathy classification. The proposed system delivers reliable diagnostic predictions while providing meaningful visual explanations that align with established ophthalmological assessment practices. By combining efficient deep learning models with explainability mechanisms, the framework enables clinicians to associate model decisions with specific retinal regions and pathological features, thereby improving transparency and trust in automated screening systems.The experimental results demonstrate that the framework performs consistently across different retinal image conditions and maintains a strong balance between accuracy and computational efficiency. The integration of lightweight architectures ensures suitability for real-time and resource-constrained environments, while the inclusion of interpretability enhances its practical usability in clinical workflows. Notably, the approach shows effectiveness in identifying early-stage diabetic retinopathy, where subtle features are often difficult to detect through conventional methods. This study contributes toward the development of scalable, transparent, and reliable AI-based solutions for diabetic retinopathy screening. The proposed framework provides a practical pathway for integrating explainable artificial intelligence into real-world healthcare systems, supporting improved clinical decision-making and better patient outcomes.

V. References

[1] Schwarcz, R. The kynurenine pathway of tryptophan degradation as a drug target. Curr. Opin. Pharmacol. 2004, 4, 12–17. [CrossRef] [PubMed]

[2] Andrzejewska-Buczko, J.; Pawlak, D.; Tankiewicz, A.; Matys, T.; Buczko, W. Possible involvement of kynurenamines in the pathogenesis of cataract in diabetic patients. Med. Sci. Monit. 2001, 7, CR742–CR745

[3] Fiedorowicz, M.; Choragiewicz, T.; Thaler, S.; Schuettauf, F.; Nowakowska, D.; Wojtunik, K.; Reibaldi, M.; Avitabile, T.; Kocki, T.; Turski, W.A. Tryptophan and kynurenine pathway metabolites in animal models of retinal and optic nerve damage: Different dynamics of changes. Front. Physiol. 2019, 10, 1254. [CrossRef] [PubMed]

[4] Kong, L.; Sun, Y.; Sun, H.; Zhang, A.-H.; Zhang, B.; Ge, N.; Wang, X.-J. Chinmedomics strategy for elucidating the pharmacological effects and discovering bio active compounds from keluoxin against diabetic retinopathy. Front. Pharmacol. 2022, 13, 728256. [CrossRef]

[5] V. S and V. R, "A Survey on Diabetic Retinopathy Disease Detection and Classification using Deep Learning Techniques," 2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII), 2021, pp. 1-4, doi: 10.1109/ICBSII51839.2021.9445163.

[6] Adak, C.; Karkera, T.; Chattopadhyay, S.; Saqib, M. Detecting Severity of Diabetic Retinopathy from Fundus Images using Ensembled Transformers. arXiv 2023, arXiv:2301.00973.

[7] Wu, J.; Hu, R.; Xiao, Z.; Chen, J.; Liu, J. Vision Transformer-

based recognition of diabetic retinopathy grade. Med. Phys. 2021, 48, 7850–7863. [CrossRef] [PubMed]

[8] Chetoui, M.; Akhloufi, M.A. Federated Learning for Diabetic Retinopathy Detection Using Vision Transformers. BioMedInformatics 2023, 3, 948–961. [CrossRef]

[9] Mohan, N.J.; Murugan, R.; Goel, T.; Roy, P. ViT-DR: Vision Transformers in Diabetic Retinopathy Grading Using Fundus Images. In Proceedings of the 2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC), Hyderabad, India, 16–18 September 2022; pp. 167–172. [CrossRef]

[10] Nazih, W.; Aseeri, A.; Youssef Atallah, O.; El-Sappagh, S. Vision Transformer Model for Predicting the Severity of Diabetic Retinopathy in Fundus Photography-Based Retina Images. IEEE Access 2023, 11, 117546–117561. [CrossRef]

[11] Kumar, N.S.; Ramaswamy Karthikeyan, B. Diabetic Retinopathy Detection using CNN, Transformer and MLP based Architectures. In Proceedings of the 2021 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Hualien City, Taiwan, 16–19 November 2021; pp. 1–2. [CrossRef]