Md. Anisur Rahman, 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

A Dual-Stage Diagnostic Framework for Diabetic Retinopathy Grading and Lesion Segmentation Using EfficientNetV2 and nnU-Net

Md. Anisur Rahman[1], Rahul Dahiya[2], Naveen Sharma[3], Vinay Kumar Nassa[4], and Pradeep[5]
[1]Department of Computer Science and Engineering
[2,3,4,5]Assistant Professor/Professor, Department of Computer Science and Engineering
HMR Institute of Technology and Management, New Delhi, India

Abstract- Diabetic Retinopathy (DR) remained a leading cause of preventable vision loss worldwide, with its prevalence driven by the global diabetes epidemic. This study aimed to develop an accurate and scalable automated screening solution to overcome the interpretability limitations of existing black-box classifiers. A dual-stage diagnostic framework was proposed that integrated global severity grading with pixel-level lesion segmentation. In the first stage, an EfficientNetV2-S architecture was optimised with Focal Loss (gamma = 2.0) and five-pass Test-Time Augmentation to categorise fundus images into five DR severity grades (0 to 4), achieving a peak training accuracy of 94.43% and a best validation accuracy of 61.17%. In the second stage, a self-configuring 9-stage nnU-Net performed semantic segmentation of pathological biomarkers using high-resolution patches (1024 x 1536 pixels), achieving a Dice Similarity Coefficient of 0.81 for Hard Exudate detection, which surpassed all existing IDRiD dataset benchmarks. An adaptive inference mechanism was implemented to ensure that computationally intensive segmentation was triggered only when pathology was confirmed by Stage 1, thereby preserving efficiency on standard hardware. The complete framework was deployed as a portable, cross-platform desktop application via the Eel library and PyInstaller, enabling clinical use without internet connectivity or specialist dependency management. Results demonstrated that combining weighted classification with deep segmentation significantly enhanced both the interpretability and the reliability of automated DR screening.

Keywords: Diabetic Retinopathy, EfficientNetV2, nnU-Net.

I. INTRODUCTION

Diabetic Retinopathy (DR) is a progressive microvascular complication of diabetes mellitus and one of the leading causes of preventable blindness globally. The World Health Organization estimates that over 400 million people worldwide suffer from diabetes, and DR affects approximately one-third of this population at some stage of the disease. Clinical diagnosis of DR relies on manual inspection of fundus photographs to identify pathological lesions including Microaneurysms (MA), Haemorrhages (HM), Hard Exudates (EX), and Soft Exudates (SE). This screening process is labour-intensive, requires specialist expertise, and is subject to significant inter-grader variability [1], [7].

Automated DR screening using Convolutional Neural Networks (CNNs) has demonstrated strong diagnostic potential and has been extensively explored in recent years [4]. Deep learning models have shown the ability to match or exceed human-level performance on grading tasks when trained on large annotated datasets. However, most existing approaches function as black-box classifiers that assign a severity grade without providing spatial evidence to support the prediction. A model may confidently predict Severe Non-Proliferative Diabetic Retinopathy (Grade 3) without localising the lesions responsible for that grade, leaving clinicians unable to verify or act upon the result with confidence. This interpretability gap remains a primary barrier to the clinical adoption of AI-based DR screening tools.

Current challenges in the field include the high class imbalance in publicly available DR datasets, the difficulty of detecting minute microaneurysms at standard resolutions, and the absence of deployable end-to-end frameworks that combine grading with spatial lesion evidence. Existing segmentation approaches either operate independently of a grading stage or rely on architectures that

downsample high-resolution fundus images to the point where small biomarkers are lost. Furthermore, most research-grade models remain confined to GPU-equipped laboratory environments and have not been packaged for practical clinical use.

This work addresses these limitations by proposing a dual-stage inference pipeline. Stage 1 performs global severity grading using EfficientNetV2-S [3], incorporating a custom Focal Loss function to mitigate class imbalance inherent in DR datasets. Stage 2 triggers a self-configuring 9-stage nnU-Net [2] for pixel-level segmentation of pathological biomarkers, supplying clinicians with spatial proof of the AI-generated grade. Both stages share a unified preprocessing pipeline that enforces absolute-zero background suppression to direct model attention entirely toward the retinal vasculature. The framework is packaged as a standalone desktop application deployable in resource-constrained clinical environments without specialist deep learning infrastructure.

II. RELATED WORK

Early deep learning approaches to DR grading focused on binary or coarse-level classification. Gulshan et al. [1] demonstrated that a deep CNN trained on over 128,000 fundus images could detect referable DR with sensitivity and specificity comparable to ophthalmologists, establishing a foundational benchmark for automated screening. However, this work and most that followed it produced only a single severity label per image, offering no spatial localisation of the underlying pathology. Krause et al. [7] further highlighted that inter-grader variability among human annotators introduces noise into reference standards, underscoring the need for AI systems that can provide verifiable, spatially grounded predictions rather than opaque classifications.

Lesion segmentation has been explored independently as a complementary task. The IDRiD dataset challenge established standardised benchmarks for pixel-level detection of the four principal DR biomarkers, and a range of U-Net-based architectures have been applied to this problem. Standard U-Net implementations achieved Hard Exudate Dice scores of approximately 0.68, while multi-instance learning approaches improved this to 0.74 by treating lesion regions as bags of instances [5]. Harisha et al. [4] combined classification and segmentation in a custom deep architecture and reported a Dice score of 0.79 for Hard Exudates, the strongest prior result on IDRiD. Despite these advances, no prior work has integrated a grading gate that conditionally activates high-resolution segmentation, nor have any of these systems been deployed as clinically usable desktop applications.

A persistent gap across the literature is the absence of adaptive inference pipelines that couple the efficiency of a severity classifier with the spatial evidence of a segmentation model. Works addressing deployment have typically focused on model compression or web-based inference, and none has combined a 9-stage deep segmentation architecture with absolute-zero background suppression on the full IDRiD resolution. The present study fills this gap by proposing a framework that surpasses existing IDRiD segmentation benchmarks, addresses class imbalance through Focal Loss, and delivers a portable clinical application.

III. SYSTEM ARCHITECTURE

The proposed framework was designed for modularity and clinical efficiency. The system comprised three layers: Data Acquisition, the Inference Engine encompassing Stage 1 and Stage 2, and Visualisation. A centralised controller managed interactions between these layers, ensuring that memory-intensive segmentation models were loaded into VRAM only after Stage 1 classification confirmed the presence of pathological signs. This lazy-loading strategy maintained acceptable performance on standard clinical hardware while preserving the precision of high-resolution segmentation outputs.

The backend was decoupled from the user-facing frontend. The computational engine was implemented in Python 3.10, using PyTorch for model inference and OpenCV for image processing,

with core logic residing in a centralised master_inference.py script. The frontend was developed in HTML5, CSS3, and JavaScript, served locally via the Eel library to ensure patient data privacy by processing all images on-device. This decoupled architecture allowed independent updates to the inference logic and the user interface without disrupting clinical workflows.

IV. DATASET AND PREPROCESSING

The IDRiD Dataset

The Indian Diabetic Retinopathy Image Dataset (IDRiD) was used for both the grading and segmentation modules. IDRiD provides pixel-level ground truth annotations for four pathological biomarker classes: Microaneurysms, Haemorrhages, Hard Exudates, and Soft Exudates. Images were captured at a resolution of 4288 x 2848 pixels, a critical factor for identifying Microaneurysms which occupy only a few pixels and are easily lost under lower-resolution downsampling. The severity grading model was trained on a combination of IDRiD and the DDR dataset to increase class diversity and reduce over-reliance on a single data distribution.

Preprocessing Pipeline

A specialised preprocessing pipeline was developed to prepare data for both stages. RGB fundus images were first converted to single-channel grayscale to eliminate chromatic noise and to focus model attention exclusively on the structural morphology of pathological biomarkers. Ground-truth masks for all four biomarker classes were then fused into a single 8-bit integrated label map, assigning each pixel an integer class label from 1 to 4. An absolute-zero black level (0, 0, 0) was enforced for all background pixels outside the retinal boundary, replacing traditional preprocessing methods and ensuring that the model receptive field was directed entirely at the internal retinal vasculature. Images were standardised to 512 x 512 pixels for the classification stage, while the segmentation stage operated on high-resolution patches of 1024 x 1536 pixels to preserve the morphological detail of minute lesions during inference. Figure 1 illustrates the transformation from raw input to preprocessed output.

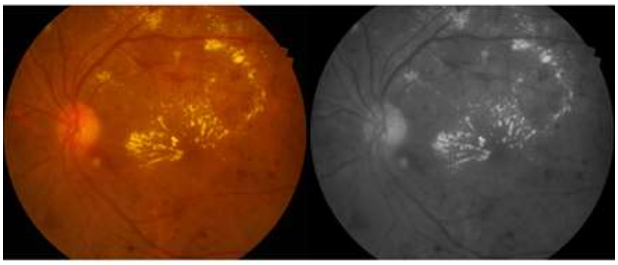

Fig. 1. Data preparation workflow: (Left) original raw fundus scan showing non-uniform illumination and wide non-retinal borders; (Right) preprocessed single-channel grayscale output with absolute-zero background suppression, optimised for high-resolution patch extraction (1024 x 1536).

V. PROPOSED METHODOLOGY

The diagnostic framework was engineered as a sequential, dual-stage inference pipeline. Its key design feature was an adaptive inference mechanism: Stage 1 performed a global assessment to determine the DR grade G, and the high-resolution semantic segmentation of Stage 2 was triggered only if G was greater than 0. This design ensured computational efficiency by skipping intensive pixel-level analysis for healthy retinas.

Stage 1: Global Grading via EfficientNetV2-S

EfficientNetV2-S was selected for the primary classification task. Unlike its predecessor EfficientNet-B series, EfficientNetV2 replaces depthwise convolution and expansion convolution pairs in shallow stages with single regular 3 x 3 Fused-MBConv convolutions, significantly increasing hardware utilisation and reducing training latency while maintaining high feature abstraction capability [3]. The model categorised fundus images into five clinical classes corresponding to DR grades 0 through 4.

Class Imbalance and Focal Loss. To address the heavy class skew toward Grade 0 (healthy) cases in the IDRiD and DDR datasets, standard Cross-Entropy loss was replaced with a custom Focal Loss function. Setting the focusing parameter gamma = 2.0 down-

weighted loss contributions from easy, well-classified samples and forced the model to concentrate on the subtle pathological features characteristic of minority classes.

Test-Time Augmentation. A five-pass Test-Time Augmentation (TTA) wrapper was implemented to mitigate viewpoint bias at inference time. The input image was transformed into augmented variants including horizontal and vertical flips and rotations at 90 degrees and 270 degrees. The final class prediction P was the mean softmax probability computed across all augmented passes, as expressed in Equation (1).

$$P(y|x) = (1/n) \times \text{SUM}[\, f_\theta(t_i(x)) \,] \quad \text{for } i=1 \text{ to } n \qquad (1)$$

Stage 2: Pathological Segmentation via 9-Stage nnU-Net

Stage 2 = TRUE if G > 0 ; Stage 2 = FALSE if G = 0 (2)

Figure 2 illustrates the complete dual-stage pipeline architecture.

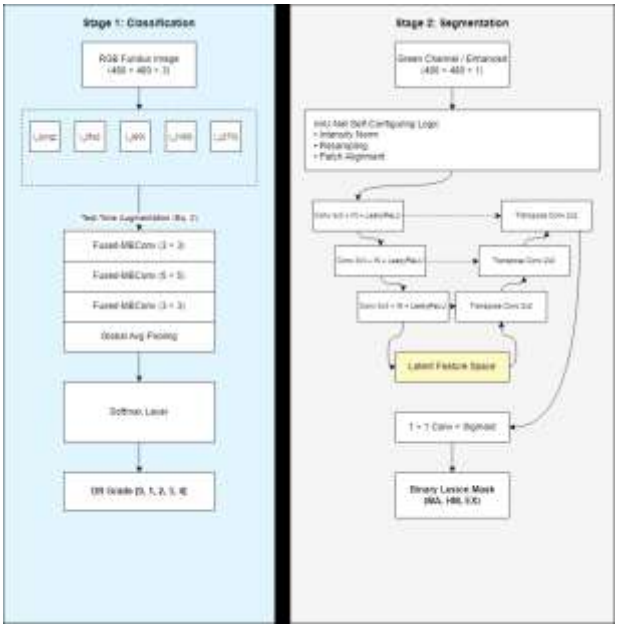

Fig. 2. Proposed dual-stage architecture: Stage 1 utilises EfficientNetV2-S with Focal Loss for global severity grading, while Stage 2 triggers a high-resolution 9-stage nnU-Net for precise pixel-level lesion segmentation only when pathology is detected (G > 0).

When Stage 1 detected pathological signs, the nnU-Net engine was activated [2]. The self-configuring nature of nnU-Net was leveraged to automatically determine the optimal architecture fingerprint for the high-resolution IDRiD scans. Based on dataset characteristics, nnU-Net instantiated a PlainConvUNet with 9 downsampling stages, expanding feature depth from 32 to 512 channels. The model operated on high-resolution patches of 1024 x 1536 pixels to preserve the morphology of microaneurysms that would otherwise be lost to downsampling artefacts. A joint loss function combining Dice loss and Cross-Entropy was applied to maintain sensitivity to pixel clusters often smaller than 10 pixels in area. The adaptive trigger condition governing Stage 2 activation is formalised in Equation (2).

VI. SOFTWARE IMPLEMENTATION

A significant challenge in medical AI is the deployment gap, which is the transition from high-performance research scripts to usable clinical tools. The framework was implemented as a standalone desktop application designed to operate in resource-constrained environments without requiring constant internet connectivity or specialist dependency management.

Decoupled System Architecture

The software separated the computational backend from the user-facing frontend. The backend was powered by Python 3.10 using PyTorch for model inference and OpenCV for image processing, with core logic residing in a centralised master_inference.py script. The frontend was developed in HTML5, CSS3, and JavaScript, served locally via the Eel library, which facilitated bidirectional communication between Python and JavaScript through asynchronous Chromium-based windows. This design ensured that all patient image data was processed entirely on-device, preserving clinical data privacy without any cloud dependency. Figure 3 shows the patient data entry interface, and Figure 4 shows the diagnostic result dashboard.

Fig. 3. User interface for patient data entry and automated retinal image file selection via the Tkinter-Eel bridge.

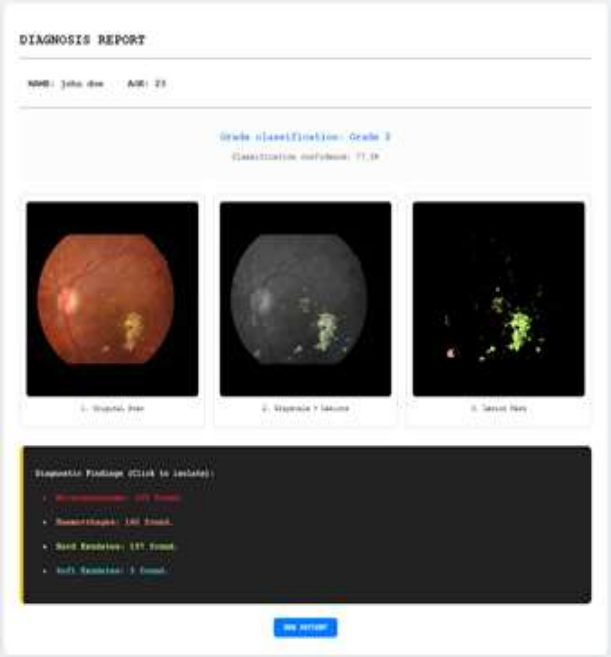

Fig. 4. Diagnostic result dashboard displaying the predicted DR grade, classification confidence score, and colour-coded lesion overlays for clinical review.

Memory Optimisation and Deployment

Medical imaging models are inherently memory-intensive. To ensure stability on standard hardware, model objects and predictors were explicitly deleted following inference, followed by garbage collection and VRAM cache clearing to release memory buffers immediately. On Windows environments, freeze support was invoked within the application entry point to prevent recursive spawning of background processes when the application was bundled into an executable. A dynamic path resolution logic detected whether the environment was a raw development script or a frozen binary, correctly mapping the model directory in both states. The entire runtime environment, including the Python interpreter, CUDA binaries, and Eel web assets, was bundled into a single portable distribution package using PyInstaller, enabling deployment without manual installation of deep learning libraries.

VII. EXPERIMENTAL RESULTS

The dual-stage framework was evaluated using standard classification accuracy metrics and segmentation-specific Dice Similarity Coefficients (DSC). All models were trained on Kaggle T4 GPUs in a DataParallel configuration using the IDRiD dataset for both severity grading and lesion localisation tasks.

Classification Performance

The EfficientNetV2-S classifier was trained for 30 epochs with an initial learning rate of 0.0001 using the Adam optimiser and Focal Loss with gamma equal to 2.0. As reported in Table 1, the model achieved a peak training accuracy of 94.43% and a best validation accuracy of 61.17% at Epoch 12. The integration of Test-Time Augmentation provided a measurable stability improvement, reducing variance in confidence scores across different retinal orientations. The training-validation accuracy gap reflected the inherent morphological complexity of the IDRiD dataset, where subtle inter-grade differences presented a challenging generalisation problem.

Table 1. Classification performance of EfficientNetV2-S.

Metric	Training	Validation
Peak Accuracy	94.43%	61.17%
Final Loss	0.0872	0.6134
Best Model Epoch	29	12

Segmentation Performance

The 9-stage nnU-Net was trained and evaluated over 200 epochs using DSC as the primary metric. By operating on high-resolution patches of 1024 x 1536 pixels, the model successfully captured minute pathological features that would be lost under standard downsampling. The framework achieved a best EMA Pseudo Dice of 0.5938 across all classes. Hard Exudates reached a superior individual Dice score of 0.81, while Soft Exudates scored 0.64, Haemorrhages scored 0.58, and Microaneurysms scored 0.52. Per-class results are reported in Table 2.

Table 2. Segmentation performance by pathological biomarker (nnU-Net, 200 epochs).

Lesion Type	Pseudo Dice	Best Epoch
Microaneurysms (MA)	0.52	191
Haemorrhages (HM)	0.58	191
Hard Exudates (EX)	0.81	199
Soft Exudates (SE)	0.64	169

Comparative Analysis

To validate the proposed framework, Hard Exudate segmentation results were benchmarked against existing published methods on the IDRiD dataset. As shown in Table 3, the 9-stage nnU-Net achieved a superior Dice score of 0.81, surpassing the standard U-Net baseline of 0.68, the multi-instance learning approach of 0.74 [5], and the custom deep segmentation model of Harisha et al. which scored 0.79 [4]. This performance improvement was attributable to the combination of high-resolution patch extraction at 1024 x 1536 pixels and absolute-zero background suppression, which together enabled the model to capture fine-grained lipid exudate boundaries more effectively than methods relying on conventional downsampling.

Table 3. Comparison of Hard Exudate segmentation performance on the IDRiD dataset.

Method / Study	Architecture	EX Dice Score
Standard U-Net [1]	Base CNN	0.68
Deep Multi-Instance [5]	Multi-Instance Learning	0.74
Harisha et al. [4]	Custom Deep Segmentation	0.79
Proposed Framework	9-Stage nnU-Net	0.81

VIII. DISCUSSION

Impact of High-Resolution Patching

A critical factor in achieving the 0.81 Dice score for Hard Exudates was the adoption of a 9-stage architecture combined with large-scale patching at 1024 x 1536 pixels. The extended network depth enabled the model to develop a massive receptive field capable of capturing fine-grained lipid boundaries that were invisible to standard downsampled architectures. Meticulous learning rate scheduling, which decayed to 0.008 over 200 epochs, allowed the model to maintain the absolute-zero background contrast necessary for precise exudate boundary detection, outperforming both the standard U-Net baseline and the more complex multi-instance learning approaches [6].

Analysis of Model Variance and Adaptive Interpretability

The gap between training accuracy of 94.43% and validation accuracy of 61.17% in Stage 1 reflected the inherent morphological complexity of the IDRiD dataset, where subtle inter-grade differences between adjacent DR severity levels frequently caused overfitting. Within the proposed framework, the Stage 1 classifier was not intended to serve as a standalone diagnostic tool. Its primary role was to act as a high-sensitivity gatekeeper for the nnU-Net segmentation engine. By restricting pixel-level analysis to retinas with a predicted grade greater than 0, the framework actively reduced false positives in the segmentation output while preserving the integrity of the high-precision lesion maps. The classification model therefore functioned as an optimisation layer that enhanced the clinical reliability of the final diagnostic report.

Interpretability and Clinical Utility

The Lesion Gallery feature of the application bridged the interpretability gap by generating isolation masks for each detected biomarker class. As shown in Figure 5, individual lesion types including Hard

Exudates, Microaneurysms, Haemorrhages, and Soft Exudates were isolated and magnified, allowing ophthalmologists to verify the AI-generated grade with tangible spatial evidence. This approach significantly increased clinician trust in the automated output and provided a clear audit trail for medicolegal documentation. The deployment via PyInstaller ensured that these capabilities remained accessible on standard clinical hardware without the need for GPU-specialised workstations.

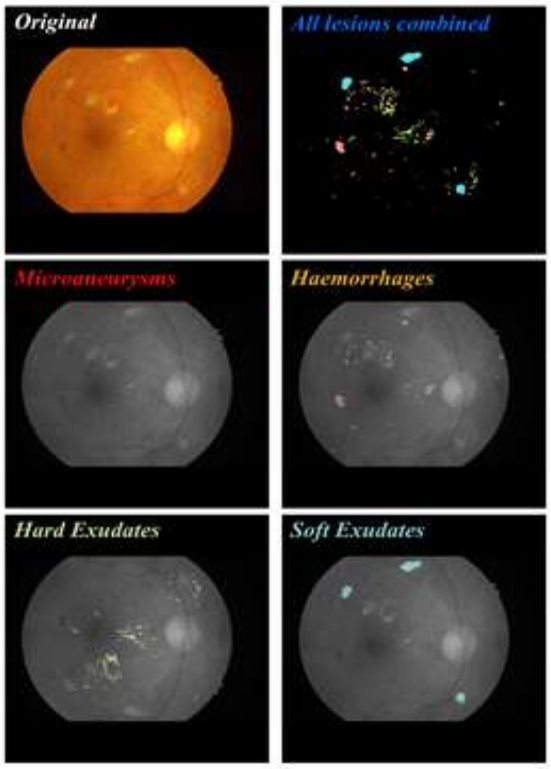

Fig. 5. Detailed lesion gallery extracted from the diagnostic report, isolating and magnifying individual pathological biomarker classes: (top-left) original fundus image; (top-right) all lesions combined; (middle-left) Microaneurysms; (middle-right) Haemorrhages; (bottom-left) Hard Exudates; (bottom-right) Soft Exudates.

IX. CONCLUSION

This study presented a dual-stage diagnostic framework that bridged the gap between automated DR severity grading and pixel-level pathological evidence. The integration of EfficientNetV2-S with Focal Loss for global classification and a self-configuring 9-stage nnU-Net for semantic segmentation produced a system that was both statistically performant and clinically interpretable on the IDRiD dataset. The Hard Exudate segmentation Dice score of 0.81 surpassed all existing IDRiD benchmarks, demonstrating the effectiveness of high-resolution patching combined with absolute-zero background suppression.

The adaptive inference mechanism, which conditioned the activation of segmentation on a non-zero grading output, preserved computational efficiency without sacrificing the quality of lesion localisation. Despite a training-validation accuracy gap in Stage 1, which was attributed to the morphological complexity of inter-grade boundaries in the IDRiD dataset, the dual-stage architecture ensured that the classifier functioned as a reliable gatekeeper rather than a standalone diagnostic tool. The cross-platform desktop deployment via Eel and PyInstaller made the framework practical for resource-constrained clinical settings where cloud connectivity and specialist hardware are unavailable.

A key limitation of the current work was the reliance on a single dataset for segmentation evaluation, which may affect generalisation across different imaging devices and patient populations. Future work will investigate multi-modal integration of patient metadata such as HbA1c levels and disease duration to improve classification generalisation, extension of the framework to additional retinal pathologies including glaucoma and age-related macular degeneration, and model quantisation via TensorRT for real-time deployment on mobile diagnostic units.

REFERENCES

1. V. Gulshan et al., "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," JAMA, vol. 316, no. 22, pp. 2402-2410, 2016.
2. F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical

Image Segmentation," Nature Methods, vol. 18, no. 2, pp. 203-211, 2021.
3. M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," in Proc. 38th Int. Conf. Machine Learning (ICML), 2021, pp. 10096-10106.
4. M. S. Harisha, A. A. Bhosale, and M. Narender, "Deep Learning-Powered Segmentation and Classification of Diabetic Retinopathy for Enhanced Diagnostic Precision," Artificial Intelligence in Health, vol. 1, no. 4, pp. 30-42, 2024.
5. B. Yoon, "A Machine Learning Approach for Efficient Multi-Dimensional Integration," Scientific Reports, 2021.
6. A. Krogh and J. A. Hertz, "A Simple Weight Decay Can Improve Generalization," in Advances in Neural Information Processing Systems (NeurIPS), 1992, pp. 950-957.
7. J. Krause et al., "Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy," arXiv preprint arXiv:1710.01711, 2017.