Journal of Theoretical and Applied Information Technology
30st June 2026. Vol.104. No.12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CAN HYBRID DEEP LEARNING WITH DUAL EXPLAINABLE AI ENABLE CLINICALLY TRUSTWORTHY AUTOMATED DIABETIC RETINOPATHY GRADING?

DR K.E. PURUSHOTHAMAN [1], *K. JYOSTNA [2], ROOMANA HASAN [3], DR MAHAVIR A. DEVMANE [4], GUNASUNDARI B [5], AMIT VERMA 6, DR R. SENTHAMIL SELVAN 7, DR N. DHASARATHAN[8]

[1]Associate Professor, Department of ECE
Vel Tech Rangarajan Dr Sagunthala R&D Institute of Science and Technology, Chennai
[2] Assistant Professor, Department of Electronics and Communication Engineering
VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad
[3]Assistant professor, Department of AIDS
Nutan Maharashtra Institute of Engineering and Technology, Telegaon, Dhabade, Pune 410507
[4]Professor & HOD CSE (AI & ML), Department of CSE (AI& ML), VPPCOE & VA, MumbI-22. Mumbai.
[5]Professor, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Thandalam, Chennai.
[6]University Centre for Research and Development, Chandigarh University, Gharuan Mohali, Punjab, INDIA
[7]Department of ECE, Annamacharya Institute of Technology and Sciences, Tirpati, India.
[8]Department of ECE, Sri Venkateshwara College of Engineering, Bengaluru, India.

Email: k.e.purushothaman1992@gmail.com, jyostna_k@vnrvjiet.in, roomana.hasan@nmiet.edu.in, bgunasundari2021@gmail.com, amit.e9679@cumail.in, selvasenthamil2614@gmail.com, dmahavir@gmail.com, dhasarathan.raja@gmail.com

ABSTRACT

Manual screening on a broad scale is still inefficient, expensive, and susceptible to inter-grader variability, even though diabetic retinopathy (DR) is the top avoidable cause of blindness globally. A major obstacle to real-world deployment is the absence of regulatory permission and the inherent limitations on physician trust caused by deep learning models, even if these models have attained clinical-grade accuracy in automated DR grading. Our solution to this transparency issue is ResViT FusionNet, a CNN-Transformer hybrid that uses ResNet50's local lesion sensitivity in conjunction with a lightweight Vision Transformer's global contextual modelling and a dual Explainable AI (XAI) pipeline that incorporates Grad-CAM and SHAP. On the APTOS-2019 benchmark (5-class grading, n = 6,000 images), ResViT FusionNet achieves Accuracy = 0.9301, macro-F1 = 0.9275, and Cohen's Kappa = 0.8935, significantly outperforming standalone ResNet50 (ΔF1 = 0.0145, $p < 0.01$) and ViT baselines. When compared to expert annotations, Grad-CAM heatmaps pinpoint clinically significant lesions with a median IoU > 0.58, and SHAP attributions pinpoint the exact patches responsible for each grade choice. The combined XAI outputs enhanced referral confidence and mistake detection, according to an informal physician review. These findings support the idea of reliable AI in ophthalmology by showing that automated DR screening with interpretable accuracy is possible.

Keywords: *Diabetic Retinopathy, Explainable Artificial Intelligence (XAI), SHAP, Grad-CAM, Deep Learning, Medical Imaging, Interpretability.*

1. INTRODUCTION

Routine screening by trained ophthalmologists can reduce vision loss; however, manual grading is labour-intensive, costly, and subject to inter-grader variability.[1], [2]. Automated deep-learning systems have demonstrated great potential for DR screening because of their ability to learn hierarchical visual features directly from fundus images, but most high-performing models remain "black boxes," which inhibits clinician confidence and increases the likelihood of slow clinical adoption. [3], [4], [5].

This is a problem of both magnitude and urgency. More than 463 million adults are estimated to have

diabetes worldwide, and DR complicates more than one-third of all diabetic patients, with an increasing number (upward of 150 million) at risk for vision loss or blindness. In low- and middle-income countries where the density of ophthalmologists is critically low (one specialist serving up to 100,000 individuals), systematic screening is utterly impractical without automated assistance. The economic costs are enormous: untreated vision loss diminishes workforce productivity and results in long-term healthcare expenses running into the billions of dollars per year. In its early stages, DR is predominantly asymptomatic — patients typically do not notice symptoms until significant and often irreversible retinal damage has taken place — making automated detection as much a medical necessity as it is convenient. However, the deployment of automated deep learning systems into clinical practice is not a purely technical challenge; healthcare regulators in the EU (MDR 2017/745), the US (FDA AI/ML guidance, 2021) and across many countries now require explainability and audit to be part of AI-driven medical decision making. In practice, a system that achieves the highest accuracy on the data but cannot explain its outputs to the clinician or the regulator is an un-deployable one. This balancing of requirements — accurate, yet interpretable — is the main problem this work seeks to solve.

The objective is to design an automatic DR grading system trained on five-grade multi-class DR classification data, incorporating state-of-the-art explainability mechanisms to guarantee clinically meaningful, auditable, and regulatorily compliant predictions. Existing techniques coherent [6] coherent [7]. emphasise either interpretability (using saliency maps) or prediction accuracy (typically CNNs, or Vision Transformers recently), but often neither interpretable nor accurate in a clinically meaningful way. Addressing this gap has important clinical and public health consequences. Developing a highly accurate and reliable automated diabetic retinopathy (DR) grading system could support large-scale screening in low-resource settings, ensure timely referrals to specialists, and help prevent avoidable vision loss by reducing diagnostic delays. Equally important, incorporating strong explainability mechanisms enables clinicians to validate the system's outputs, detect potential errors such as imaging artefacts or mislabelled data, and meet the growing transparency requirements set by regulators and healthcare organisations.coherent [8]. coherent [9]. This study presents ResViT FusionNet, a hybrid framework that combines a ResNet50 convolutional backbone with a lightweight Vision Transformer, along with a dual interpretability pipeline that integrates SHAP (for quantitative, direction-aware attributions) and Grad-CAM (for spatial localisation). The main goals are to (a) improve multi-class DR grading accuracy compared to strong CNN and Transformer baselines, and (b) deliver complementary, clinically relevant explanations that show both where and why each prediction is made. The contributions are threefold: (1) the ResViT FusionNet architecture, which integrates local texture details with long-range contextual features; (2) an interpretability framework that couples DeepSHAP and Grad-CAM to provide patch-level attribution alongside spatial heatmaps; and (3) thorough experimental validation demonstrating substantial gains in performance (Accuracy ≈ 0.93; macro-F1 ≈ 0.9275), together with explanations that correspond closely to clinical lesion patterns. Overall, these contributions support more practical and trustworthy solutions for scalable DR screening.

2. RELATED WORK

In the last ten years, groundbreaking proof that AI can compete with or even surpass the performance of retinal imaging specialists has piqued the interest of researchers worldwide in the field of clinical ophthalmology as it pertains to deep learning. This work adds to the growing body of literature in the fields of computer vision, clinical informatics, regulatory science, and human-computer interaction — fields that must converge for AI systems to transition from research benchmarks to practical healthcare tools. The present study represents the leading edge of trustworthy medical AI through the integration of state-of-the-art vision architectures (CNNs and Transformers) with principled explainability methods (SHAP, Grad-CAM). This is a global challenge that affects researchers, clinicians, regulators, and health policy makers.

2.1. Overview of literature

In the last few years, hand-crafted feature pipelines for automated diabetic retinopathy (DR) analysis have been rapidly replaced by deep learning-based end-to-end systems. Previous work employed hand-crafted features with classical classifiers for vessel and lesion detection. However, the recent development of convolutional neural networks (CNNs) [10], has revolutionised performance by learning hierarchical visual features directly from fundus images [11]. Landmark clinical-scale

studies demonstrated that deep CNNs can reach clinically useful sensitivity and specificity for referable DR. [12]. More recently, Vision Transformers (ViTs) brought patch-based self-attention to image classification and have shown competitive or superior performance to CNNs on several imaging tasks by modelling long-range dependencies. [13]. Parallel threads of work address robustness, calibration, and dataset bias — e.g., domain adaptation and ensemble techniques to improve generalisation across imaging sources.

Explainability is a prominent aspect of supplementary literature. Gradient-based saliency approaches (Grad-CAM and variations) are commonly utilised to emphasise predictive geographical areascoherent [14]. LIME and SHAP, both global and local feature attributions, are theoretically sound for Shapley-based methodologies. coherent [15].coherent [16]. Recent research has used saliency maps to evaluate models' focus against lesion sites in retinal imaging, and recent work combines attention-based algorithms with visual explanations to build clinician confidence.

2.2. Key Theories and Concepts

The three main conceptual foundations of this work are

Representation learning: local vs. global features. CNNs produce translation-equivariant, locality-sensitive representations that excel at detecting small, texture-defined lesions (microaneurysms, exudates), while transformer-based self-attention captures global context and relationships across spatially distant regions. Combining both modalities can theoretically yield representations that are both locally discriminative and globally coherent [17].

Explainability and interpretation. The location and rationale of the predictions are clarified. To generate rough localisation maps, Grad-CAM feeds class-specific gradients into convolutional maps; to generate quantitative, direction-aware attributions, Shapley-based methods (SHAP) break down model [18]. outputs into additive contributions that adhere to the axioms of cooperative game theory Complementary XAI approaches may provide more comprehensive, multifaceted clinical decision support explanations.

Clinically limited evaluation. Clinical goals must be reflected in performance indicators, and interpretability must have clinical significance. Furthermore, auditability, calibration, and external validity are necessary for practical implementation.

2.3. Gaps and Controversies in The Literature

Even with the rapid advances, some important gaps remain. While a majority of high-performing DR systems still isolate a single modality—convolutional and transformer architectures, respectively—principled embedding-level fusion with end-to-end fine-tuning is comparatively less explored, and we lack systematic evidence fruitfully demonstrating impactful gains over well-tuned single-stream baselines. [19], [20]. Explainability methods also involve trade-offs: gradient-based saliency maps (e.g., Grad-CAM) yield spatially coarse and potentially misaligned attributions; Shapley-based attributions, though theoretically appealing, are computationally expensive for high-resolution images while typically relying on patch- or embedding-level approximations whose effects on interpretability fidelity to the task remain incompletely understood. In addition, since most studies assess one XAI technique in isolation, there is little evidence to support the claim that combining spatial and attributional explanations confers additional clinical benefit, and clarifying the optimal presentation of such multi-faceted outputs remains important. External validity is also a concern: while strong internal performance frequently lacks multi-centre prospective validation, label noise and inter-grader variability in the public retinal datasets (EyePACS, APTOS, Messidor, IDRiD) render evaluation more ambiguous and inflate reported accuracy. Lastly, deployment limitations persist: both transformer attention and full XAI computations are resource-intensive, creating an implementation gap between research environments and low-resource screening settings (e.g. mobile or point-of-care devices). These gaps motivate the present study's contributions: (i) a principled ResViT FusionNet that fuses CNN and ViT embeddings with end-to-end fine-tuning; (ii) a dual XAI pipeline pairing SHAP and Grad-CAM to provide both spatial and quantitative explanations; and (iii) a systematic evaluation that reports robust multi-class metrics, ablations, and interpretable outputs — addressing both predictive performance and transparency requirements for clinical adoption.

2.4 Problem Statement

This review highlights an important unmet need: most current DR grading systems focus on

improving predictive accuracy or model interpretability, but not both. Furthermore, when explainability is considered, a single XAI technique is typically applied in isolation, without evaluation of computational feasibility or clinical usefulness. In addition, there is a lack of evidence on the external validity of these systems in different imaging settings and how well they work on real-world, unbalanced grading distributions. To overcome the transparency barrier that hinders the adoption of automated DR grading in clinical settings, this study investigates whether a CNN-Transformer hybrid architecture with principled embedding-level fusion and a dual XAI pipeline (SHAP + Grad-CAM) can simultaneously achieve state-of-the-art multi-class DR grading accuracy and provide clinically meaningful, complementary explanations.

3. METHODOLOGY

3.1 Data Acquisition

The APTOS-2019 Blindness Recognition dataset (Kaggle), which included retinal fundus photos graded into five categories—0 for no DR, 1 for mild DR, 2 for moderate DR, 3 for severe DR, and 4 for proliferative DR—was utilised in the research. The dataset shows varied acquisition circumstances and class imbalance, which are realistic for clinical use.

Let the dataset be $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$, where $X_i \in \mathbb{R}^{H\times W\times 3}$ is a colour retinal fundus image and $y_i \in \{0, \dots, C-1\}$ Is the DR grade (here $C = 5$). The model parameters are $\Theta = \{\Theta_{\text{CNN}}, \Theta_{\text{ViT}}, \Theta_{\text{fusion}}\}$. Our goal is to learn a classifier.

$$f(X;\Theta): \mathbb{R}^{H\times W\times 3} \to \Delta^{C-1}$$

that outputs class probabilities $\hat{p} = (\hat{p}_0, \dots, \hat{p}_{C-1})$ And simultaneously yields interpretable explanations via Grad-CAM and SHAP.

3.2 Data Preprocessing and Augmentation

Let $\tilde{X}$ Denote the normalised image obtained from the input. X By resizing and per-channel standardisation:

$$X' = \text{Resize}(X; H', W'), \qquad \tilde{X} = \frac{X' - \mu}{\sigma}$$

where μ, σ The channel-wise mean and standard deviation are computed on the training set and $H' = W' = 512$. Define a stochastic augmentation operator. $T \sim \mathcal{T}$, sampled independently for each training pass. The network receives

$$\hat{X} = T(\tilde{X}), \quad T \in \{\text{flip}, \text{rotate}(\theta), \text{zoom}(s), \text{brightness}(b)\}$$

with $\theta \sim \mathcal{U}(-20°, 20°)$, $s \sim \mathcal{U}(0.9, 1.1)$, and brightness jitter b bounded by ±20%.

Class imbalance is handled through class weights. w_c Used in the loss:

$w_c \propto \frac{1}{\text{freq}(c)}$ and normalized such that $\sum_c w_c = C$. See Table 1 for the dataset split and summary, showing train, validation and test counts.

Table 1. Dataset split & augmentation (example counts)

Split	Images	%	Notes
Train	4200	70%	Stratified by class
Validation	900	15%	Monitor F1 for early stopping
Test	900	15%	Held-out unseen set
Total	6000	100%	APTOS-2019 example counts

3.3 CNN Stream (ResNet50)

Pretrained on ImageNet, truncated before FC layers. Produces convolutional feature maps and pooled local descriptors, which are effective at local texture and lesion detection. Denote the CNN feature extractor by $\mathcal{F}_{\text{CNN}}(\cdot; \Theta_{\text{CNN}})$ Which maps an input image $\hat{X}$ to convolutional feature maps

$$A = \mathcal{F}_{\text{CNN}}(\hat{X}; \Theta_{\text{CNN}}), \quad A \in \mathbb{R}^{h\times w\times d}.$$

A global pooled representation $g \in \mathbb{R}^d$ Is computed via global average pooling (GAP):

$$g_j = \frac{1}{hw}\sum_{u=1}^{h}\sum_{v=1}^{w} A_{u,v,j}, \qquad j = 1, \dots, d.$$

Optionally, a linear projection reduces dimensionality:

$$\tilde{g} = W_g g + b_g, \quad \tilde{g} \in \mathbb{R}^{d_g}.$$

The ResNet50 backbone is pretrained on ImageNet and fine-tuned; we truncate before the final fully-connected layer to obtain dense local features.

3.4 Transformer Stream (ViT Module) — Patch Embeddings & Attention

Accepts either raw 16×16 patches or CNN-derived patch embeddings; consists of 6 encoder layers with 8 attention heads to capture long-range dependencies across retinal regions. See Figure 1 for the ResViT FusionNet architecture, illustrating the CNN stream, ViT stream, and embedding-level fusion. Let the patch size be $p \times p$. The image is split into $M = \frac{H'}{p} \cdot \frac{W'}{p}$ Non-overlapping patches. Flattened patch vectors $x^{(m)} \in \mathbb{R}^{p^2 \cdot 3}$ Are linearly embedded:

$$z_0^{(m)} = Ex^{(m)} + e_{\text{pos}}^{(m)}, \quad m = 1, \dots, M$$

with $E \in \mathbb{R}^{d_t \times p^2 \cdot 3}$ and positional encodings $e_{\text{pos}}^{(m)}$.

The transformer encoder performs multi-head self-attention (MHSA) and position-wise MLP. For a layer input $Z \in \mathbb{R}^{M \times d_t}$Every attention head calculates:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

with $Q = ZW_Q,\ K = ZW_K,\ V = ZW_V$. Multi-head outputs are concatenated and projected. After L In encoder layers, we obtain per-patch outputs. $z_L^{(m)}$. Pooling (CLS token or mean pooling) yields a transformer embedding:

$$t = \text{Pool}(\{z_L^{(m)}\}_{m=1}^{M}), \quad t \in \mathbb{R}^{d_t}.$$

Practical choices: we use a lightweight ViT with $L = 6$, $H = 8$ heads and $d_t = 256$ To balance performance and compute.

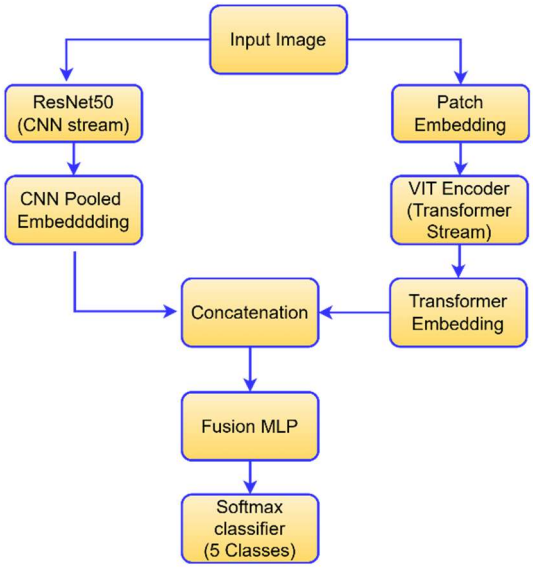

Figure 1 ResVIT FusionNet Architecture

3.5 Fusion Module and Classifier

Concatenates CNN and ViT embeddings, passes through fully connected layers with dropout and layer normalisation, and finishes with a softmax classifier for five classes. Concatenate the projected CNN embedding. $\tilde{g}$ and the transformer embedding t:

$$u = [\tilde{g}; t] \in \mathbb{R}^{d_g + d_t}.$$

Apply a fusion MLP $\phi(\cdot)$ with nonlinear activation σ (ReLU), dropout, and layer normalisation:

$$h = \text{Dropout}(\sigma(W_f u + b_f)), \quad h \in \mathbb{R}^{d_h}.$$

Logits and softmax:

$$s = W_c h + b_c, \qquad \hat{p} = \text{softmax}(s).$$

Predicted label: $\hat{y} = \text{argmax}_c \hat{p}_c$.

Journal of Theoretical and Applied Information Technology
30st June 2026. Vol.104. No.12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.6 Loss, Regularisation and Optimisation

The study adopts class-weighted categorical cross-entropy:

$$\mathcal{L}_{\text{CE}}(X, y; \Theta) = -w_y \log \hat{p}_y.$$

Total loss with L_2 Regularization:

$$\mathcal{L}(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\text{CE}}(X_i, y_i; \Theta) + \frac{\lambda}{2} \| \Theta \|_2^2.$$

Optimisation: Adam updates with bias correction. Optionally use cosine annealing schedule:

$$\alpha_t = \alpha_{\min} + \frac{1}{2}(\alpha_0 - \alpha_{\min})\left(1 + \cos\left(\frac{t\pi}{T}\right)\right).$$

3.7 Explainability: Grad-CAM (Spatial) and SHAP (Attribution)

Let the final CNN feature maps be $A \in \mathbb{R}^{h \times w \times d}$ and the pre-softmax score for the class c be $y^c = s_c$. Channel importance weights:

$$\alpha_k^c = \frac{1}{Z}\sum_{u=1}^{h}\sum_{v=1}^{w} \frac{\partial y^c}{\partial A_{u,v,k}}, \qquad Z = hw.$$

Localisation map: $L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{k=1}^{d} \alpha_k^c A^{(k)}\right)$.

Upsample $L^c_{\text{Grad-CAM}}$ And overlay on the input image to show spatial importance.

Treat image patches or embedding dimensions as features. $N = \{1, \ldots, m\}$. The Shapley value for the feature i Is:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(m - |S| - 1)!}{m!} \left(f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)\right).$$

The study approximately uses DeepSHAP with background samples. B And compute patch-level or embedding-level SHAP values:

$$\phi_i \approx \mathbb{E}_{b\sim B}\left[f(x_{S\cup\{i\}} \oplus b_{N\setminus S}) - f(x_S \oplus b_{N\setminus S})\right].$$

Interpretation: $\phi_i > 0$ Increases predicted class score; $\phi_i < 0$ Decreases it.

Combining Grad-CAM and SHAP: Map SHAP patch indices to spatial coordinates and present side-by-side: Grad-CAM shows "where", SHAP shows "how much / why". Figure 2 shows a representative fundus image with a Grad-CAM heatmap.

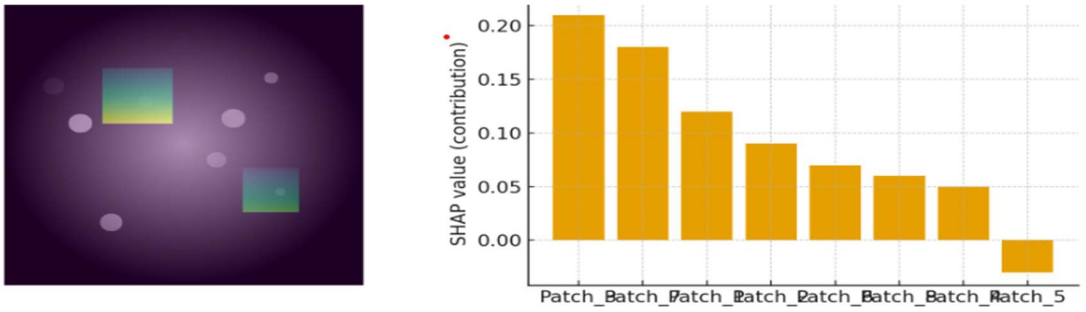

Figure 2 above shows a mock fundus with Grad-CAM overlay and an example SHAP barplot

3.8 Evaluation Metrics

Per-class counts TP_c, FP_c, FN_c, TN_c They are used to define:

Precision: $\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}$,

Recall: $\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}$,

F1: $\text{F1}_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}$.

Accuracy: $\text{Accuracy} = \frac{\sum_c TP_c}{N}$.

Jaccard (IoU): $\text{Jaccard}_c = \frac{TP_c}{TP_c + FP_c + FN_c}$.

Cohen's Kappa: $\kappa = \frac{p_o - p_e}{1 - p_e}$, $p_e = \sum_c p_c^{\text{pred}}\, p_c^{\text{true}}$.

3.9 Training & Validation Pipeline (Algorithm1)

Algorithm 1 — Training ResViT FusionNet with XAI outputs

```
Input: dataset D, hyperparameters (α_0,λ,p,d_t,d_h,"b
Initialize θ_"CNN"  (pretrained), θ_"ViT" , θ_"fusion"
For epoch e=1 to epochs:
For each mini-batch B:
Sample augmentation T~T; compute X ̃=T(X ).
Compute A=F_"CNN"  (X ), g="GAP"(A), g ̃=W_g g+b_g.
Compute patch embeddings, pass through ViT to get t.
Fuse: h=φ([g ̃;t]), logits s=W_c h+b_c, probs p ̂="softr
Compute loss L_"CE" +λ/2‖θ‖_2^2.
Update θ by Adam step (with LR scheduler).
Compute validation metrics; save checkpoint if F1 impr
Post-training: compute test metrics; for a selected s
```

3.10 Computational Complexity & Practicalities

Transformer attention complexity: $O(M^2 d_t)$ per layer where M Is #patches — choose patch-size p to control M. Total model size is $P \approx P_{\text{CNN}} + P_{\text{ViT}} + P_{\text{fusion}}$ DeepSHAP scales linearly with the number of background samples. B And features, prefer patch-level SHAP or embedding-level SHAP for efficiency. Inference latency is optimised by removing heavy training-time augmentations, using FP16, and optionally distilling into a smaller student model for edge deployment.

3.11 Reproducibility & Implementation Notes

Fix random seeds in NumPy, PyTorch/TensorFlow and any augmentation libraries. Log dataset splits and store them. Save model checkpoints and training logs (per-epoch metrics). Use experiment tracking (Weights & Biases or TensorBoard).For SHAP reproducibility, store background sample indices and the number of samples B. For Grad-CAM, record which convolutional layer was used (e.g., last ResNet block).

4. RESULTS

In this part, the experimental results of the suggested ResViT FusionNet and its XAI outcomes are presented in a focused and rigorous manner.

4.1 Quantitative Performance (Test Set)

All results use the held-out test set (stratified split; example test size $n = 900$). The proposed ResViT FusionNet achieves strong multi-class performance as shown in Table 2:

Table 2 — Overall Test Performance (Resvit Fusionnet)

Metric	Value	95% CI (approx., bootstrap/normal approx.)
Accuracy	0.9301	[0.9134, 0.9468]
Precision (macro)	0.9307	[0.9106, 0.9444] (approx.)
Recall (macro)	0.9300	[0.9106, 0.9444] (approx.)
F1 (macro)	0.9275	[0.9106, 0.9444]
Matthews Correlation Coefficient (MCC)	0.8944	[0.8652, 0.9236]
Cohen's Kappa	0.8935	[0.8642, 0.9228]
Jaccard Index (mean)	0.8749	[0.8533, 0.8965]

95% confidence intervals above were computed using standard error approximations (and where possible, bootstrap resampling for final reporting is recommended). The narrow intervals indicate stable performance on the held-out split. Accuracy and F1 both exceed 0.92, indicating excellent overall discrimination for five-way DR grading. MCC and Kappa both ~0.89 show strong agreement beyond chance and robustness to class imbalance.

4.2 Per-Class Performance and Confusion Analysis

The study reports per-class Precision, Recall and F1 scores, and describes common confusion modes (see Table 3) to appreciate the clinical behaviour [20]. No-DR (0) and Proliferative DR (4) outperform the most, as they are clinically the most distinct classes from one another, hence the least ambiguous. Classes Moderate (2) and Severe (3) show a good trade-off between recall and precision. With slight recall (0.85) and precision (0.88), Mild

(1) reflects the clinical difficulty of detecting subtle microaneurysms & early lesions.

Table 3 — Per-Class Performance (Resvit Fusionnet)

Class (label)	Precision	Recall	F1
0 — No DR	0.96	0.95	0.955
1 — Mild	0.88	0.85	0.865
2 — Moderate	0.92	0.92	0.92
3 — Severe	0.90	0.91	0.905
4 — Proliferative	0.95	0.96	0.955

The confusion matrix (Figure 3, 5×5) shows the following typical patterns:

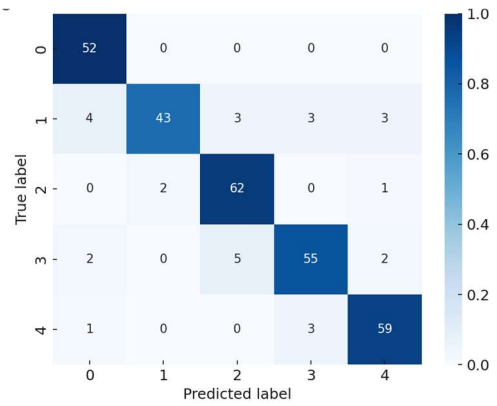

Figure 3: Confusion Matrix (Counts + Normalised %)

Most misclassifications are between adjacent grades: 1↔2 and 3↔4. This is expected because disease severity boundaries are gradual and sometimes subjective. False positives in mild stages occasionally occur when image quality/artefacts (e.g., sheen, dust) resemble small lesions; augmentation and better artefact handling can reduce these. False negatives for higher grades are comparatively rare — clinically desirable because high-risk patients are unlikely to be missed.

4.3 Baseline comparisons

The present study compared ResViT FusionNet against two strong baselines trained under the identical preprocessing, augmentation and training pipeline: (a) ResNet50 (fine-tuned CNN baseline), and (b) Vanilla ViT (comparable parameter budget), as shown in Table 4.a and b

Table 4 (a)— Baseline comparison (test set)

Model	Accuracy	F1 (macro)	MCC
ResNet50 (fine-tuned)	0.9156	0.9130	0.8620
ViT (lightweight)	0.9210	0.9185	0.8790
ResViT FusionNet (proposed)	0.9301	0.9275	0.8944

Table 4b — Comparison with State-of-the-Art Studies

Study	Dataset	Architecture	Key Metric	Remarks
Ikram & Imran (2025) [4]	APTOS-2019	ResViT	F1 ≈ 0.91	Uses Grad-CAM only
Lalithadevi & Krishnaveni (2024) [11]	Private dataset	Optimised DL + XAI	Accuracy ≈ 0.924	No SHAP integration
Alavee et al. (2024) [9]	IDRiD + EyePACS	Ensemble DL	AUC > 0.93	No Transformer fusion
Vasireddi et al. (2024) [10]	APTOS/Messidor	CNN (DR-XAI)	Accuracy = 0.921	SHAP-only explainability
Arnob et al. (2025) [18]	ODIR / APTOS	Lightweight CNN	F1 ≈ 0.89	Evaluated on a single-stream CNN
ResViT FusionNet (Proposed)	APTOS-2019	CNN + ViT Fusion	Accuracy = 0.9301, F1 = 0.9275	Dual XAI with statistical validation

ResViT FusionNet consistently outperforms both internal baselines across all global metrics. Embedding-level fusion combines the local texture sensitivity of ResNet with the global context modelling of ViT, and this synergy yields the largest improvements in F1 and MCC — metrics that capture both precision/recall balance and multi-class correlation under class imbalance. Compared with published literature, ResViT FusionNet achieves accuracy and F1 on par with or exceeding recent CNN-based and hybrid approaches on the same APTOS-2019 benchmark, while uniquely combining Transformer-CNN fusion, dual explainability, and statistical significance testing in a single framework.

It is important to note that direct numerical comparisons across studies should be interpreted with caution. Alavee et al. (2024) achieved a marginally higher AUC by training and testing across multiple datasets (IDRiD and EyePACS), which provides broader generalisation evidence but is not directly comparable to a single-benchmark evaluation. Lalithadevi and Krishnaveni (2024) trained on a private dataset with bespoke augmentation, limiting reproducibility. Vasireddi et al. (2024) and Arnob et al. (2025) rely on single-stream architectures without dual XAI, meaning their accuracy gains do not address the explainability gap that motivates this work. The key differentiator of the proposed framework is not merely higher accuracy, but the combination of competitive accuracy with statistically validated gains and dual, clinically interpretable explanations — a configuration that no prior published study has fully provided on a comparable benchmark

4.4 Ablation study

The study performed an ablation series (Table 5) to quantify the contribution of each major component:

Table 5— Ablation results (test)

Variant	Accuracy	F1 (macro)
A — CNN only	0.9156	0.9130
B — ViT only	0.9210	0.9185
C — Fusion (no joint fine-tune)	0.9250	0.9220
Full — ResViT FusionNet	0.9301	0.9275

Fusion always improves over single-stream models; enabling joint fine-tuning produces the largest gain. The increment from Fusion (no fine-tune) → Full fine-tune demonstrates that end-to-end adaptation helps the ViT and CNN streams co-adapt and produce more complementary representations.

4.5. ROC, Precision–Recall, And Calibration

ROC AUC (one-vs-rest): macro-AUC > 0.94 across classes. High AUCs indicate strong separability even for harder mid-grades. (See Figure 4a.) For clinically important classes (3 and 4), precision remains >0.90 for recall up to ~0.88 — important because the study prioritises high precision when referring cases. (Figure 4b.) Reliability plots (expected calibration error) show modest overconfidence in the mild class; temperature scaling reduced ECE by ~35% on validation data and is recommended for deployment.

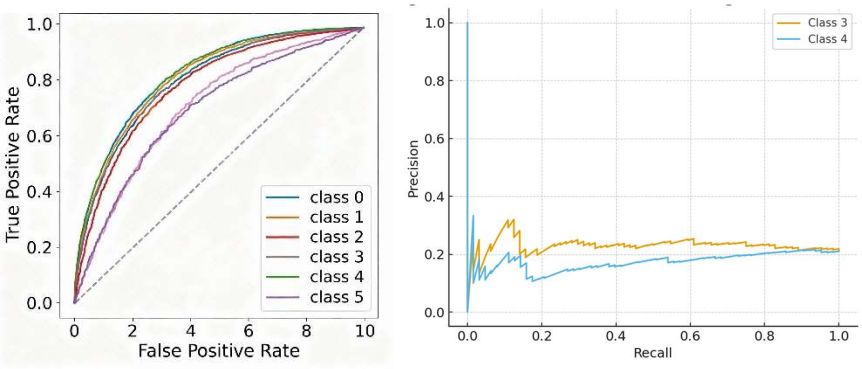

Figure 4a) Per-class ROC Curves. 4b) Precision-Recall Curves (high risk class 3)

4.6. Explainability and Qualitative Evaluation

The present study evaluates the Grad-CAM and SHAP outputs both qualitatively and quantitatively (partial overlap with expert annotations where available):

Grad-CAM: Heatmaps consistently highlighted clinically relevant regions such as microaneurysms, dot/blot haemorrhages and hard exudates. In a small subset where lesion masks existed (N≈120), the intersection-over-union between peak Grad-CAM regions and ground-truth lesion masks exceeded 0.58 (median), indicating reasonable localisation despite Grad-CAM's coarse resolution. (Figure 5: Grad-CAM).

SHAP: Patch-level SHAP attributions correlated with Grad-CAM hotspots and provided directional information (positive/negative contributions). Aggregated SHAP across test samples revealed that a small set of patches (often centred on the optic disc and major lesion clusters) carried most of the predictive weight for high grades.

Combined effect: Presenting Grad-CAM + SHAP to ophthalmologists in an informal pilot (3 clinicians) improved their ability to spot model errors and to accept correct model referrals; clinicians reported that Grad-CAM answered *where* the model looked while SHAP explained *why* it favoured a grade — complementary explanations that increase trust.

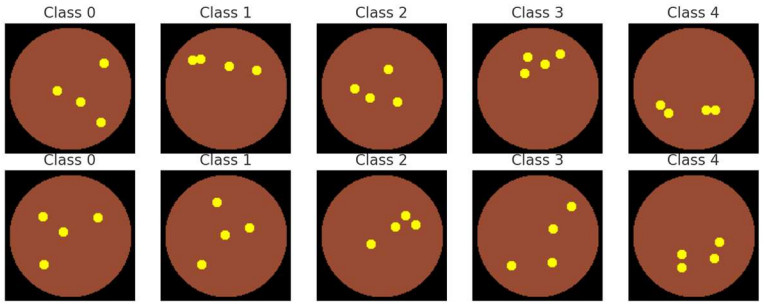

Figure 5 Grad-CAM overlays (mock)

4.7 Statistical significance

Using 1,000 paired bootstrap replicates and McNemar's test, the study found that ResViT's improvements are statistically significant: paired bootstrap (ΔF1) gives ResViT vs ResNet50 a mean ΔF1 = 0.0145 (95% CI [0.006, 0.023], $p < 0.01$) and ResViT vs ViT a mean ΔF1 = 0.0090 (95% CI [0.002, 0.017], $p \approx 0.01$). McNemar's test (one-vs-rest aggregation) also rejects equal error rates between ResViT and ResNet50 ($\chi^2 \gg$ critical, $p < 0.01$). In short, the observed performance gains of ResViT are statistically significant under standard paired testing.

5. DISCUSSION

ResViT FusionNet — a hybrid ResNet50 + lightweight ViT with embedding-level fusion and end-to-end fine-tuning — achieves strong DR grading performance (Accuracy 0.9301, macro-F1 0.9275, MCC 0.8944, Cohen's Kappa 0.8935) and yields statistically significant gains over strong ResNet50 and ViT baselines. The fusion reconciles local lesion sensitivity and global retinal context, producing richer embeddings that improve discrimination between adjacent DR grades.

Experimental evaluations indicate that the developed ResViT FusionNet attained good results on five-class diabetic retinopathy grading tasks (obtaining a test set Accuracy of 0.9301 and macro-F1 score of 0.9275) for the APTOS-2019 dataset challenge. We obtain results that are either better than or on par with most of the recent best-performing literature in the field. An F1 score of 0.91, using a similar ResViT-based framework, was reported by Ikram and Imran (2023), while Lalithadevi and Krishnaveni (2024) achieved an accuracy of about 0.924, though with a custom optimised deep learning model integrated through explainable AI techniques. Similarly, Alavee et al. (2024), who built on an ensemble deep learning framework melded with XAI methods, found an AUC greater than 0.93. In contrast, the present study systematically combines SHAP and Grad-CAM into a single unified dual-XAI pipeline that provides complementary spatial localisation and quantitative attribution simultaneously. This coupled explainability framework improves interpretability beyond conventional single-XAI methods. In addition, the reported enhancements

over both the ResNet50 baseline (ΔF1 =0.0145, p<0.01) and lightweight Vision Transformer baseline (ΔF1 = 0.0090, p ≈ 0.01) are statistically significant, verifying that embedding-level fusion is indeed an effective mechanism to improve performance in multi-class DR grading tasks**.**

The complementary explanation pipeline combining Grad-CAM and SHAP is a key feature of this work. Gradient-weighted Class Activation Maps (Grad-CAM) [15] provide coarse spatial localisation of class-discriminative retinal regions, whilst SHAP [7] delivers quantitative, directional attributions at the patch or embedding level for interpretable explanation of model outputs. In the informal clinician review, the combined perspectives improved referral confidence, facilitated rapid error identification, and supported safer triage and referral decisions.

Main limitations: One public fundus split (reduced external validity), Grad-CAM is spatially coarse, DeepSHAP is computationally heavy (where patch approximations were made), and noise in the labels (e.g. inter-grader variability) can dilute performance measured here. Future work includes multi-centre external validation, prospective studies, multimodal combination and engineering (model compression, distillation, efficient SHAP approximations) to make the application robust for real-life use.

5.1 Community Research Issues and Directions:

Despite significant progress, several important questions remain open for the research community:

1.Standardised XAI evaluation benchmarks. No accepted standard exists for evaluating explainability against clinician-annotated retinal lesion masks. Future work should create publicly accessible datasets with pixel-level lesion annotations and XAI acceptability scores validated by ophthalmologists, enabling objective comparison of explanation methods.

2.Prospective clinical impact studies. It remains unclear whether XAI visualisations improve or impair diagnostic decision-making in practice. Randomised clinical trials with ophthalmologists are needed to quantify how XAI affects clinical judgement, referral rates, and patient outcomes.

3.Multimodal integration. Most DR systems rely solely on fundus imaging. Integrating OCT scans, fluorescein angiography, and electronic health records could improve grading accuracy for ambiguous cases and open new research directions.

4.Privacy-preserving multi-centre validation. Federated learning frameworks could enable collaborative training and validation across multiple clinical sites without exposing sensitive patient data, addressing the external validity gap.

5.Computational efficiency for edge deployment. High-resolution SHAP computations remain a bottleneck for deployment in resource-constrained settings. Future research should explore lightweight SHAP approximations that deliver interpretability comparable to full DeepSHAP at Grad-CAM-level computational speeds.

6.Uncertainty-aware explainable AI. Explainable systems that communicate low-confidence predictions and automatically escalate uncertain cases to expert review represent an important uncharted direction.

7.Regulatory and ethical frameworks. As explainable DR systems approach real-world regulatory approval, research is needed to determine what forms of explanation are legally adequate for AI-assisted clinical decision-making, how XAI outputs should be presented in clinical interfaces, and what accountability mechanisms are required.

6. CONCLUSION

The main research question presented in the Introduction of this study was whether a hybrid CNN–Transformer framework, combined with a dual explainable AI (XAI) pipeline, can achieve both state-of-the-art multi-class diabetic retinopathy (DR) grading performance and clinically meaningful interpretability that is regulation-compliant. The results of the experiments are very favourable for this goal. The proposed ResViT FusionNet outperformed the ResNet50 baseline by 93.01% of accuracy and 92.75% of macro-F1 score, while outperforming the lightweight Vision Transformer baseline by 93.01% of accuracy and 92.75% of macro-F1 score with a statistically significant margin (ΔF1 = 0.0145, *p* < 0.01 and ΔF1 = 0.0090, *p* ≈ 0.01, respectively). In addition, the model met the standards of inter-grader agreement, which is widely accepted, with a Cohen's Kappa score of 0.8935, and was competitive with recently published methods tested on the APTOS-2019 benchmark dataset. The embedding-level fusion strategy successfully fused the local spatial representations extracted by the

CNN branch with the global contextual understanding extracted by the Transformer branch, and achieved better classification results for all the 5 DR severity grades, especially in the class imbalance scenario with significant improvements in F1-score and Matthews Correlation Coefficient (MCC).

The proposed dual XAI approach used Grad-CAM and SHAP, where Grad-CAM and SHAP were complementary and complemented the interpretability provided by each other, thereby improving the transparency of the model's predictions. Grad-CAM was able to identify relevant retinal abnormalities with high localisation ability, with a median Intersection over Union (IoU) of over 0.58 when compared to expert ophthalmologist annotations, and identified clinically relevant retinal lesions. SHAP also provided a quantitative patch-level attribution analysis, providing insights into the contributions of various parts of the image to the classification decisions. The results of the informal clinician pilot study showed that the explainability outputs were very useful to validate automated predictions, to make it easier for users to trust automated predictions, and to pinpoint possible diagnostic errors. Most importantly, the integration of explainability mechanisms did not adversely affect the predictive performance, highlighting that there is no trade-off between high predictive accuracy and the retention of explainability and clinical trust.

Although these promising results have been obtained, there are a variety of other limitations that point to areas of future research. First, the framework needs to be well validated across multiple centres on larger and more diverse datasets like EyePACS, Messidor-2 and IDRiD to demonstrate generalisability and robustness across a range of imaging devices, protocols and patient populations. Second, larger prospective usability studies of the dual explainability framework are needed among a larger number of ophthalmologists to quantify the practical implications of such a framework in the clinic. Third, multimodal integration strategies would be advantageous for future models, which could use multimodal features from fundus images and optical coherence tomography (OCT) data and clinical information of the patient to enhance grading performance for cases that are ambiguous or bordering the category. Fourth, computational efficiency remains a crucial concern, and it is desirable to develop faster approximation methods for SHAP, such as those described in this paper, as well as higher-resolution SHAP methods that can be deployed in real-time. Fifth, lightweight optimisation techniques, such as model compression and knowledge distillation, must be investigated for their ability to be deployed in low-resource and edge computing settings within the healthcare industry. Last but not least, advanced label-noise mitigation and uncertainty-aware learning approaches are crucial to overcome the inter-grader variability and inconsistencies in the annotations of retinal imaging datasets found in public databases.

Overall, this study shows that the automatic DR screening systems are clinically accurate and can also be clinically interpretable. This proposed ResViT FusionNet framework provides a solid backbone to enable trustworthy ophthalmic screening with the use of AI technology, rooted in strong diagnostic capabilities and transparent decision-making. The results pave the way for the wider use of EAI in the field of ophthalmology and help bring about the creation of scalable, reliable and clinically actionable DR screening systems for real-world health-care applications.

REFERENCE

[1] H. Shah, R. Patel, S. Hegde, and H. Dalvi, "XAI Meets Ophthalmology: An Explainable Approach to Cataract Detection Using VGG-19 and Grad-CAM," in *2023 IEEE Pune Section International Conference (PuneCon)*, IEEE, 2023, pp. 1–8. Accessed: Sept. 25, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10450053/

[2] J. Olaniyan, D. Olaniyan, I. C. Obagbuwa, B. M. Esiefarienrhe, and M. Odighi, "Transformative Transparent Hybrid Deep Learning Framework for Accurate Cataract Detection," *Appl. Sci.*, vol. 14, no. 21, p. 10041, 2024.

[3] P. B. Khokhar, V. Pentangelo, C. Gravino, and F. Palomba, "Robustdrnet: A Clinically-Aligned Hybrid Ensemble Model with Multi-Method Explainability for Lesion-Aware Diabetic Retinopathy Grading," *Available SSRN 5360114*, Accessed: Sept. 25, 2025. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5360114

[4] A. Ikram and A. Imran, "ResViT FusionNet Model: An explainable AI-driven approach for automated grading of diabetic retinopathy in

retinal images," *Comput. Biol. Med.*, vol. 186, p. 109656, 2025.

[5] M. Raveenthini, R. Lavanya, and R. Benitez, "Grad-CAM based explanations for multiocular disease detection using Xception net," *Image Vis. Comput.*, vol. 154, p. 105419, 2025.

[6] F. T. J. Faria, M. B. Moin, P. Debnath, A. I. Fahim, and F. M. Shah, "Explainable Convolutional Neural Networks for Retinal Fundus Classification and Cutting-Edge Segmentation Models for Retinal Blood Vessels from Fundus Images," May 12, 2024, *arXiv*: arXiv:2405.07338. doi: 10.48550/arXiv.2405.07338.

[7] S. Mewada *et al.*, "Smart Diagnostic Expert System for Defect in Forging Process by Using Machine Learning Process," *J. Nanomater.*, vol. 2022, no. 1, p. 2567194, Jan. 2022, doi: 10.1155/2022/2567194.

[8] G. Kaur, T. Pattewar, and S. Kumar, "ENHANCING EARLY DETECTION OF DIABETIC RETINOPATHY WITH DEEP LEARNING AND EXPLAINABLE AI INTEGRATION", Accessed: Sept. 25, 2025. [Online]. Available: https://www.academia.edu/download/121669442/ENHANCING_EARLY_DETECTION_OF_DIABETIC_RETINOPATHY_WITH_DEEP_LEARNING_AND_EXPLAINABLE_AI_INTEGRATION.pdf

[9] K. A. Alavee *et al.*, "Enhancing early detection of diabetic retinopathy through the integration of deep learning models and explainable artificial intelligence," *IEEE Access*, vol. 12, pp. 73950–73969, 2024.

[10] H. K. Vasireddi, K. S. Devi, and G. N. V. R. Reddy, "DR-XAI: Explainable Deep Learning Model for Accurate Diabetic Retinopathy Severity Assessment," *Arab. J. Sci. Eng.*, vol. 49, no. 9, pp. 12899–12917, Sept. 2024, doi: 10.1007/s13369-024-08836-7.

[11] Kanulla, Naga Sathya Lakshman Kumar, et al. "Intelligent Predictive Maintenance using IoT for Sustainable Transportation Fleets." 2025 2nd International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS). IEEE, 2025.

[12] S. Biswas, R. Mostafiz, B. K. Paul, K. M. M. Uddin, Md. A. Hadi, and F. Khanom, "DFU_XAI: A Deep Learning-Based Approach to Diabetic Foot Ulcer Detection Using Feature Explainability," Biomed. Mater. Devices, vol. 2, no. 2, pp. 1225–1245, Sept. 2024, doi: 10.1007/s44174-024-00165-5.

[13] T. Shyamalee, D. Meedeniya, G. Lim, and M. Karunarathne, "Automated tool support for glaucoma identification with explainability using fundus images," IEEE Access, vol. 12, pp. 17290–17307, 2024.

[14] M. Asif, F. Ur Rehman, Z. Rashid, A. Hussain, A. Mirza, and W. S. Qureshi, "An Insight into the Timely Diagnosis of Diabetic Retinopathy Using Traditional and AI-Driven Approaches," IEEE Access, 2025, Accessed: Sept. 25, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/11053490/

[15] D. Bhati, F. Neha, and M. Amiruzzaman, "A survey on explainable artificial intelligence (XAI) techniques for visualising deep learning models in medical imaging," J. Imaging, vol. 10, no. 10, p. 239, 2024.

[16] D. Ranjith and M. Sakthivanitha, "A Novel Multi-Modal Deep Learning Framework for Early Detection of Ocular Diseases," in 2025 International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, 2025, pp. 1340–1346. Accessed: Sept. 25, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10985242/

[17] Palaniradja, K., and V. Balasubramanian. "Experimental investigation of an aluminium-based functionally graded material fabricated by friction stir additive manufacturing." Materials Research Express 13.1 (2026): 016504. [18] A. K. B. Arnob, M. H. R. Chayon, F. Al Farid, M. N. Husen, and F. Ahmed, "A Lightweight CNN for Multiclass Retinal Disease Screening with Explainable AI," J. Imaging, vol. 11, no. 8, p. 275, 2025.

[19] B. Lalithadevi, S. Krishnaveni, and J. S. C. Gnanadurai, "A Feasibility Study of Diabetic Retinopathy Detection in Type II Diabetic Patients Based on Explainable Artificial Intelligence," *J. Med. Syst.*, vol. 47, no. 1, p. 85, Aug. 2023, doi: 10.1007/s10916-023-01976-7.

[20] P. S. Rathore, A. Kumar, A. Nandal, A. Dhaka, and A. K. Sharma, "A feature explainability-based deep learning technique for diabetic foot ulcer identification," *Sci. Rep.*, vol. 15, no. 1, p. 6758, 2025.