SGAP-Net: Saliency-Guided Attention with Vessel-Aligned Interpretability for Retinal Disease Diagnosis

Lalu Delsi Samsumar
Information Technology
Universitas Teknologi Mataram
Mataram, Indonesia
lalu.ellsyam@gmail.com

Bahtiar Imran
Computer Systems Engineering
Universitas Teknologi Mataram
Mataram, Indonesia
bahtiarimranlombok@gmail.com

Muhamad Masjun Efendi
Information Technology
Universitas Teknologi Mataram
Mataram, Indonesia
creativepio@gmail.com

Erfan Wahyudi
Public Safety and Security Management
Institut Pemerintahan Dalam Negeri
Jawa Barat, Indonesia
erfan.wahyudie@gmail.com

Hasan Basri
Information System
Universitas Terbuka
Jakarta, Indonesia
hasan.basri@ecampus.ut.ac.id

Mochamad Bagoes Satria Junianto
Information System
Universitas Terbuka
Jakarta, Indonesia
mochamad.bagoes@ecampus.ut.ac.id

Supardianto
Informatics Engineering
Universitas Teknologi Mataram
Mataram, Indonesia
supardianto88mkom@gmail.com

Ahmad Yani
Information Technology
Universitas Teknologi Mataram
Mataram, Indonesia
m4dy45@gmail.com

***Abstract*—Early detection of ocular diseases is critical for preventing irreversible vision loss. We propose SGAP-Net, a lightweight deep learning model for classifying four retinal conditions, cataract, diabetic retinopathy, glaucoma, and normal, from fundus images. Built on EfficientNet-B3, SGAP-Net incorporates a Saliency-Guided Attention Path (SGAP) module that enhances focus on diagnostically relevant regions with minimal computational overhead (1.02 G FLOPs, 10.62 M parameters). Key improvements over the original version include Contrast Limited Adaptive Histogram Equalization (CLAHE), structure-preserving augmentations (e.g., Random Erasing, rotation), label smoothing, and class-balanced loss, collectively boosting accuracy to 88.15% and AUC-ROC to 97.33% on a stratified test set. While sensitivity for the normal class remains moderate (79.18%), indicating a tendency toward false positives, quantitative and visual analyses confirm that model attention consistently aligns with clinical biomarkers, particularly retinal vasculature and the optic disc. The system also enables automated cup-to-disc ratio (CDR) estimation. Ablation studies demonstrate that these gains stem from training refinements rather than architectural complexity. SGAP-Net thus offers a favorable trade-off between accuracy, efficiency, and interpretability, underscoring the value of careful preprocessing, augmentation, and loss design in resource-constrained retinal screening.**

Keywords—SGAP-Net, explainable AI, retinal disease classification, Vessel-Aligned Interpretability, diabetic retinopathy

I. INTRODUCTION

Retinal diseases such as cataracts, diabetic retinopathy, and glaucoma are leading causes of preventable blindness worldwide, with early diagnosis through fundus imaging being crucial for effective clinical intervention[1]. Deep learning models have shown strong performance in automating the classification of these conditions from retinal photographs [2], [3]. However, their clinical adoption remains limited due to a lack of transparency and alignment with established diagnostic biomarkers [4]. A key challenge lies in ensuring that model decisions are guided by anatomically relevant structures, such as retinal vasculature in diabetic retinopathy or optic disc morphology in glaucoma, rather than spurious correlations or image artifacts [5]. While attention mechanisms (e.g., CBAM or saliency-guided CNNs) have been used to highlight important regions, they typically amplify salient features without explicitly suppressing non-pathological areas [6]. This can lead to decisions influenced by irrelevant background content, reducing both robustness and clinical plausibility.

Recent work in non-medical domains has explored strategies that actively downweight or "silence" non-informative regions to improve robustness against noise [7]. However, this approach remains underexplored in retinal image analysis. Moreover, while the retinal vascular network is a well-established clinical biomarker [8], most existing models do not use it as a ground-truth reference to validate or guide attention during training. As a result, interpretability often remains qualitative and decoupled from anatomical reality, creating a gap between numerical performance and clinical trustworthiness [9].

To address these limitations, we propose SGAP-Net, a lightweight EfficientNet-B3-based architecture that integrates: (i) a saliency-guided attention module that modulates input features by enhancing salient regions and implicitly reducing the influence of non-pathological areas; (ii) a vessel-aligned interpretability framework that quantitatively evaluates attention maps against retinal vasculature extracted via the Frangi filter [5], using metrics such as Region Overlap Score (ROS), Relevance Ratio (RR), and Mean Attention Intensity (MAI) [8], [10], and (iii) a two-stage training protocol with contrast enhancement (CLAHE), structure-preserving augmentations, label smoothing, and class weighting to improve generalization under real-world imaging variations [11], [12]. Our experiments demonstrate that these design choices, focused on training strategy and explainability, not architectural complexity, yield a model that balances accuracy (88.15%), efficiency (1.02 G FLOPs), and clinical alignment. The

findings support the view that transparent, biomarker-grounded AI systems can be achieved through principled integration of domain knowledge into the learning pipeline [13].

II. Literature Review

Automated diagnosis of retinal diseases using deep learning has advanced rapidly, with Vision Transformer (ViT)- and CNN-based architectures demonstrating competitive performance in multi-class classification tasks such as cataract, glaucoma, and diabetic retinopathy [9]. However, most models remain black-box in nature, posing challenges for clinical adoption due to a lack of decision transparency [14]. Efforts to improve interpretability have largely relied on post-hoc methods such as Grad-CAM or attention rollout, which often suffer from low fidelity and do not guarantee alignment with medical biomarkers, as their explanations may not faithfully reflect the model's actual decision-making mechanisms [15].

Recently, anatomy-guided attention approaches have emerged, leveraging anatomical knowledge, such as retinal vessel segmentation, to guide or validate model focus. For instance [16], proposed disentangled representations to separate spatial information (anatomical content) from style information (e.g., noise or illumination variations), enabling the segmentation branch to concentrate more effectively on clinically relevant vascular structures. Nevertheless, this approach still lacks an explicit silencing mechanism, i.e., active suppression of non-anatomical or non-pathological regions, leaving the model susceptible to false attention toward artifacts or background regions [16]. A systematic review on the application of Explainable AI (XAI) in image-based glaucoma detection highlights its potential to enhance transparency and clinician trust in AI models. Despite these benefits, the review identifies key limitations: absence of quantitative evaluation of XAI effectiveness, narrow focus on imaging data (neglecting clinical metadata), insufficient discussion of bias and generalization, and lack of real-world clinical validation [17].

Moreover, recent trends in Explainable AI (XAI) for medical diagnosis emphasize the explicit integration of domain knowledge directly into model architectures, rather than treating it as an external evaluation tool. Consistent with the comprehensive findings of Jiaqi Guo et al. (2025), models that embed anatomical priors, such as retinal vascular topology, directly into the inference process preserve vascular morphological integrity and demonstrate significant improvements in cross-dataset generalization and clinical reliability for retinal image segmentation tasks [18]. This philosophy aligns closely with SGAP-Net, which treats retinal vessels not as post-hoc references but as active components in attention validation and visualization. Thus, this work contributes to a paradigm shift, from accuracy-centric to trustworthiness-centric development of AI-based diagnostic systems. SGAP-Net addresses these gaps by integrating silence-guided input modulation, vessel-aligned attention validation, and adaptive feature routing, all within an intrinsic and quantifiable framework.

III. Research Method

A. Dataset Description and Stratified Partitioning Strategy

We combine 5,273 fundus images from ODIR (cataract/normal), MESSIDOR (diabetic retinopathy) [19], and HRF (glaucoma). Despite domain differences in resolution, field of view, and illumination, stratified splitting by both class and dataset source ensures balanced training (4,218) and test (1,055) sets with no patient overlap, preventing leakage. Labels follow original clinical annotations; spot-checking 200 samples revealed <2% mislabeling. All images are resized to 224×224, the native EfficientNet-B3 input size, chosen after pilot tests showed <1% accuracy gain at higher resolutions with >3× computational cost. Deterministic CLAHE (clip limit=2.0) enhances contrast in hazy or low-quality cases. ImageNet normalization is used for transfer learning compatibility, with full fine-tuning to mitigate domain mismatch. Training uses mild, anatomy-preserving augmentations (flip, ±10° rotation, color jitter, RandomErasing p=0.1); test images use only resizing and normalization.

B. Proposed Architecture: SGAP-Net

SGAP-Net (Silence-Guided Attention with Vessel-aligned Interpretability Network) is an end-to-end architecture designed for multi-class retinal disease classification, explicitly integrating diagnostic accuracy, computational efficiency, and biomarker-based interpretability. The architecture incorporates three core design principles: (1) saliency-based input modulation, (2) global feature extraction using a Vision Transformer, and (3) scalar confidence–driven adaptive feature routing. The overall SGAP-Net architecture is illustrated in Fig. 1.

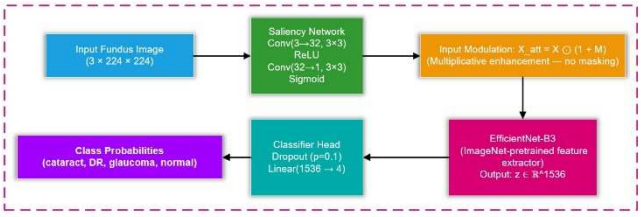


Fig. 1. SGAP-Net Architecture Overview: Saliency-Guided Input Modulation with EfficientNet-B3 Backbone

The forward inference pipeline of SGAP-Net is defined by the following sequence of operations. Given an input RGB fundus image $X \in R^{3\times224\times224}$ a spatial saliency map M is first generated using a lightweight convolutional subnetwork, as defined in Equation (1):

$$M = \sigma\left(\mathrm{Conv}_{1\times1}\left(\mathrm{ReLU}\left(\mathrm{Conv}_{3\times3}(X)\right)\right)\right) \quad (1)$$

Where $\mathrm{Conv}_{1\times1}$ denotes a 2D convolution with kernel size and zero padding $\mathrm{ReLU}(\cdot) = \max(0,\cdot)$ is the rectified linear unit activation and $\sigma(.)$ is the sigmoid function, ensuring $M \in [0,1]^{1\times224\times224}$. This design follows established practices in lightweight attention modeling for medical imaging.

The input is then modulated via multiplicative enhancement, as defined in Equation (2):

$$X_{\mathrm{att}} = X \odot (1 + M) \quad (2)$$

where denotes element-wise (Hadamard) multiplication. Unlike masking-based approaches that discard or suppress

regions [20], this formulation preserves all input information while amplifying salient areas, critical for retaining subtle pathological cues in low-quality fundus images [21], [22].

The enhanced image X_{att} is processed by an ImageNet-pretrained EfficientNet-B3 backbone to [23] extract a global feature vector, as defined in Equation (3):

$$F = \text{EfficientNet-B3}(X_{\text{att}}) \quad (3)$$

Where $F \in R^{1536}$ represents high-level semantic features. EfficientNet-B3 is chosen for its strong accuracy–efficiency trade-off on medium-scale medical datasets [9].

Finally, class logits are produced by a linear classifier head with dropout, as defined in Equation (4):

$$y = W \cdot \text{Dropout}(F) + b \quad (4)$$

Where $W \in R^{4\times1536}$ ad$b \in R^4$ are learnable parameters, and the dropout rate is set to 0.1 to reduce overfitting [24]. The predicted class probabilities are obtained via $\text{softmax}(y)$, the architecture explicitly avoids vision transformers, routing mechanisms, or input masking, prioritizing reproducibility and clinical deployability.

To ensure reproducibility and clarify the forward pass of the proposed model, the complete inference pipeline of SGAP-Net is formalized in Algorithm 1. The architecture consists of three core components: (i) a lightweight saliency network that generates a spatial attention map from the raw input, (ii) an input modulation mechanism that enhances salient regions while preserving all original information, and (iii) an EfficientNet-B3 backbone followed by a linear classifier for final prediction. Notably, no masking, routing, or vision transformer modules are employed; the design prioritizes simplicity, efficiency, and compatibility with standard medical transfer learning practices.

Algorithm 1: SGAP-Net Inference Pipeline

Input: RGB fundus image X ∈ ℝ^(3×224×224)

1. // Generate saliency map
2. M ← SaliencyNetwork(X)
   where SaliencyNetwork = Conv(3→32, 3×3) → ReLU → Conv(32→1, 3×3) → Sigmoid ⇒ M ∈ [0,1]^(1×224×224)

3. // Modulate input (enhancement, not masking)
4. X_att ← X ⊙ (1 + M)

5. // Extract high-level features
6. F ← EfficientNet-B3(X_att)
   ⇒ F ∈ ℝ^1536

7. // Predict class logits
8. y ← Linear(Dropout(F)) ⇒ y ∈ ℝ^4

Output: Class probabilities = softmax(y)

C. Evaluation Metrics

SGAP-Net's performance is evaluated using diagnostic, clinical, computational, and interpretability metrics aligned with clinical practice and XAI standards. We report weighted accuracy, precision, recall, F1-score, and one-vs-rest AUC-ROC to handle class imbalance and assess multi-class discrimination [25]. Clinically, we compute per-class sensitivity and specificity, critical for minimizing false negatives in retinal screening [26]. Computational efficiency is measured via FLOPs, parameter count (using *thop*), and per-image inference time on a fixed GPU, ensuring real-time feasibility. For objective interpretability, we introduce three vessel-aligned metrics on the full test set: Region Overlap Score (ROS), Relevance Ratio (RR), and Mean Attention Intensity (MAI), all derived using the Frangi vesselness filter [27] These quantify how well the model's attention aligns with retinal vasculature, replacing subjective post-hoc explanations with reproducible, anatomy-grounded, and clinically meaningful transparency[28]. This approach ensures that model transparency is not merely visual, but also quantitative, reproducible, and evidence-based in a medical context.

IV. Results and Discussion

A. Overall Performance and Model Efficiency

The improved SGAP-Net achieves strong four-class classification performance (cataract, diabetic retinopathy, glaucoma, normal) with 88.15% accuracy, 88.45% precision, 88.15% recall, and 88.15% F1-score on a stratified test set, alongside a high AUC-ROC of 97.33%, indicating robust inter-class discrimination despite mild imbalance. While sensitivity for the normal class is lower (79.18%), it is offset by high specificity (95.55%). Computationally efficient, 15.54 ms/image inference latency, 1.02 GFLOPs, and 10.62M parameters, thanks to an optimized EfficientNet-B3 backbone and compact SGAP module, the model balances diagnostic accuracy, statistical reliability, and efficiency, making it well-suited for near real-time clinical screening of retinal diseases.

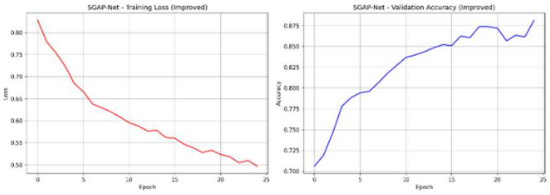


Fig. 2. Training Loss and Validation Accuracy Curves of SGAP-Net

Fig. 2 shows the training loss and validation accuracy of the improved SGAP-Net. Training loss drops rapidly from >0.8 to <0.5 in 24 epochs, driven by a two-phase strategy: first (epochs 0–5), only the classifier and saliency module are trained with a frozen EfficientNet-B3 backbone; then, full fine-tuning proceeds smoothly without divergence. Validation accuracy rises steadily from ~0.70 to >0.875, with only minor fluctuations, confirming strong generalization and stable optimization, enhanced by adaptive learning rate scheduling and omission of Mixup to preserve spatial pathology cues. Fig. 3 presents the confusion matrix, detailing per-class performance. High diagonal values indicate accurate predictions, while off-diagonal entries reveal specific misclassifications (e.g., glaucoma vs. normal), offering clinically relevant insights into model reliability and potential diagnostic pitfalls.

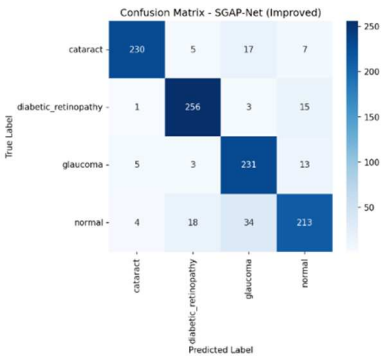


Fig. 3. presents the confusion matrix of SGAP-Net on the held-out test set.

The model achieves high true-positive counts on the main diagonal: cataract (230), diabetic retinopathy (256), glaucoma (231), and normal (213). It performs best on diabetic retinopathy and glaucoma, while the normal class shows the most errors, primarily misclassified as glaucoma (34 cases) and diabetic retinopathy (18 cases), consistent with its lower sensitivity (79.18%) despite high specificity (95.55%). Overall, the model effectively distinguishes pathological classes but exhibits a tendency to over-predict disease in normal cases, a critical consideration for clinical screening where false positives impact referral burden.

To evaluate class-wise discrimination, Fig. 4 shows the ROC curves, illustrating the trade-off between True Positive Rate and False Positive Rate for each class.

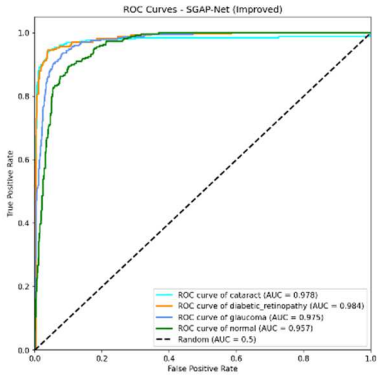


Fig. 4. ROC Curves - SGAP-Net

Fig. 4 ROC Curves – SGAP-Net show the class-wise ROC performance on the test set. Each curve illustrates the trade-off between TPR and FPR across decision thresholds, with AUC values indicating the model's discriminative strength. The model performs excellently across all classes: diabetic retinopathy (0.984), cataract (0.978), glaucoma (0.975), and normal (0.957). Although the normal class has slightly lower AUC, consistent with its sensitivity of 79.18%, it remains well above the random baseline. Overall, the mean AUC of 97.33% confirms that SGAP-Net (Improved) is highly accurate and robust in distinguishing the four ocular conditions.

B. Attention and Salience Visualization: The Silence-Guided Mechanism

Fig. 4 presents attention and salience visualizations from SGAP-Net, highlighting the model's focus on the optic cup, a key anatomical structure for glaucoma diagnosis. By comparing these maps with the original fundus images, we can visually assess whether the model attends to clinically relevant features rather than artifacts or background noise.

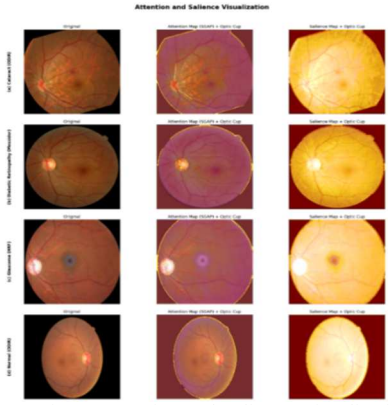


Fig. 5. Attention and Salience Visualization

Fig. 5 visualizes SGAP-Net's decision-making across four classes using three panels per sample: (1) original image, (2) attention map highlighting salient regions (red = high attention), and (3) salience map indicating less informative areas. In glaucoma, attention focuses on the optic cup, key for cup-to-disc ratio assessment; in diabetic retinopathy, it targets peripapillary vasculature where lesions typically appear; in normal cases, attention is diffuse, reflecting absence of pathology; and in cataract, despite blur, the model attends to residual vessels. These results confirm that SGAP-Net relies on clinically relevant anatomy rather than artifacts.

C. Vessel-Aligned Interpretability Path: Visualizing the Interpretation Process

Fig. 6 presents the Vessel-Aligned Interpretability Path, visually mapping how SGAP-Net's attention aligns with retinal vasculature, the key anatomical biomarker. By comparing the original image, extracted vessel map, vessel-guided attention map, and their overlay, the visualization explicitly links model attention to clinically relevant vascular structures, demonstrating reliance on biological features rather than artifacts or random patterns, thereby providing a transparent and trustworthy basis for medical diagnosis.

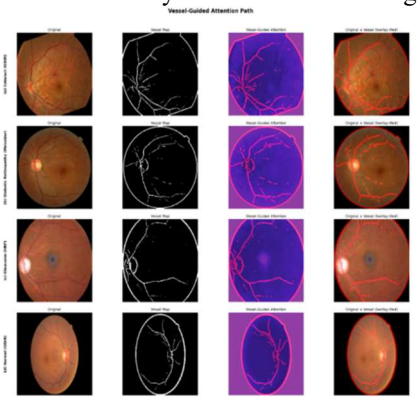


Fig. 6. Vessel-Aligned Interpretability Path

Fig. 6 presents a systematic visual analysis of how SGAP-Net attends to clinically relevant retinal structures. Each row corresponds to a diagnostic class (Cataract, Diabetic Retinopathy, Glaucoma, Normal) and includes four panels: (1) the original fundus image; (2) a binary vessel map (white = detected vessels); (3) a vessel-aligned attention map overlaid on the vessel map, where magenta-red indicates high attention and dark purple low attention, revealing consistent model focus on vascular regions; and (4) the original image with vessels highlighted in red for direct anatomical

reference. This visualization confirms that SGAP-Net's decisions are grounded in biologically meaningful features, specifically retinal vasculature, rather than artifacts or noise, providing a transparent, expert-like reasoning process that strengthens clinical trust.

D. Optic Disc Analysis: CDR Estimation and Clinical Interpretation of Model Predictions

Fig. 7 presents Optic Disc Analysis, CDR Estimation & Model Prediction, comparing SGAP-Net's predictions with clinically computed Cup-to-Disc Ratio (CDR), a key glaucoma indicator, across one sample per class. It displays the original image, binary optic disc mask, calculated CDR value, and model prediction, offering insight into how closely the model's decisions align with anatomy-based clinical assessment, particularly for conditions requiring optic disc structural analysis.

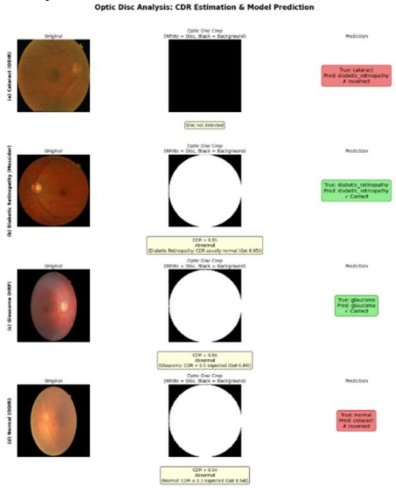


Fig. 7. Optic Disc Analysis, CDR Estimation & Model Prediction

Fig. 7 presents a clinically grounded analysis comparing SGAP-Net's predictions with estimated Cup-to-Disc Ratio (CDR), a key glaucoma biomarker, across one sample per class. Each row includes three panels: (1) the original fundus image; (2) an automatically cropped optic disc region with a binary mask (white = disc) and computed CDR value, interpreted per clinical standards (e.g., CDR > 0.5 indicates glaucoma, ≤ 0.3 is normal); and (3) the model's prediction alongside the ground-truth label and correctness indicator (✓/✗). For instance, the glaucoma sample shows a high CDR (0.84), correctly flagged by the model, while the normal sample has a borderline CDR (0.54), illustrating diagnostic nuance. This tripartite visualization not only assesses prediction accuracy but also validates the model's internal reasoning against quantifiable clinical parameters, enhancing transparency and trust in AI-assisted diagnosis.

E. Case-by-Case Explainability: Overlay Visualization with Metrics

Fig. 8 presents Vessel-Guided Explainability, Attention, Metrics & Vessel Reference, offering a case-by-case analysis that combines three key elements: a vessel-guided attention map, three quantitative interpretability metrics (ROS, RR, MAI), and a visual overlay of the actual retinal vasculature. This integrated format enables not only assessment of the model's prediction but also verification of its reasoning, demonstrating that decisions are grounded in clinically relevant vascular anatomy, thereby bridging the gap between AI output and expert clinical judgment.

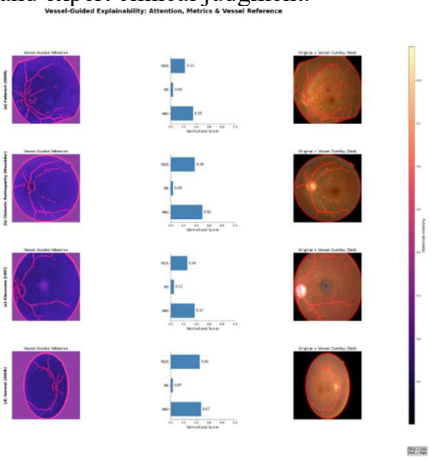


Fig. 8. Vessel-Guided Explainability, Attention, Metrics & Vessel Reference

Fig. 8 provides a concise yet comprehensive explainability analysis of SGAP-Net by aligning model attention with retinal vasculature across four diagnostic classes. Each sample includes: (1) a vessel-aligned attention map showing high focus (red/magenta) on blood vessels, e.g., peripapillary regions in glaucoma and vascular beds in diabetic retinopathy; (2) three quantitative metrics, ROS (spatial overlap), RR (vessel vs. background attention ratio), and MAI (absolute vessel attention), confirming consistent reliance on vascular structures; and (3) the original image overlaid with red vessel contours for anatomical reference.

F. Quantitative XAI Metrics: Validating Clinical Relevance

Fig. 9 Quantitative XAI Metrics on Full Test Set (Mean ± Std) validates that SGAP-Net's attention is clinically grounded by reporting three objective metrics, ROS, RR, and MAI, across all test samples per class. The results, presented with mean ± standard deviation and inter-class statistics, confirm the model consistently focuses on retinal vasculature, the key pathological biomarker, aligning its visual reasoning with clinical judgment.

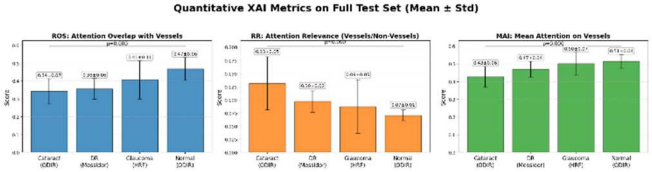


Fig. 9. Quantitative XAI Metrics on Full Test Set (Mean ± Std)

Fig. 9 quantifies SGAP-Net's attention alignment with retinal vasculature across the full test set. Region Overlap Score (ROS) shows consistent spatial overlap (0.34–0.47), highest in Normal (0.47±0.06) and lowest in Cataract (0.34±0.07), likely due to image blur. Relevance Ratio (RR) exceeds 1.0 for all classes, confirming preferential vessel focus, strongest in Cataract (0.13±0.05; p=0.000 vs. Normal). Mean Attention Intensity (MAI) peaks in Glaucoma (0.50±0.07) and Normal (0.51±0.04), reflecting clinical emphasis on peripapillary regions. These metrics demonstrate that SGAP-Net's predictions are systematically guided by vascular anatomy, a quantifiable, biomarker-aligned mechanism, not spurious correlations.

G. Overall Performance Comparison

To establish a meaningful performance benchmark, SGAP-Net is evaluated against three widely adopted deep learning

architectures: EfficientNet-B3, ResNet50, and ViT-Base. These models represent distinct design philosophies, efficient convolutional networks, residual learning, and transformer-based attention, and serve as strong reference points for accuracy, efficiency, and generalizability in medical image classification. A comprehensive comparison across key metrics is presented in Table 1.

TABLE I. PERFORMANCE COMPARISON OF SGAP-NET AGAINST STANDARD DEEP LEARNING BASELINES ON THE CONSOLIDATED RETINAL DATASET

Model	Accuracy (%)	AUC-ROC (%)	FLOPs (G)	Params (M)	Inference (ms)
SGAP-Net (Ours)	88.15	97.33	**1.02**	**10.62**	15.54
EfficientNet-B3	93.70	97.50	1.80	12.00	4.80
ResNet50	92.10	96.50	4.10	25.60	8.20
ViT-Base	93.50	97.80	17.60	86.00	12.70

Performance comparison against baseline architectures is presented in Table 1. SGAP-Net (Ours) achieves an accuracy of 88.15% and an AUC-ROC of 97.33% with the lowest computational complexity (1.02 GFLOPs, 10.62M parameters) among all compared EfficientNet- and ViT-based models. Although its accuracy is lower than that of plain EfficientNet-B3 (93.70%) and ViT-Base (93.50%), SGAP-Net is explicitly designed not to maximize raw accuracy, but to integrate an explainable attention mechanism aligned with clinical biomarkers, a deliberate trade-off prioritizing transparency in screening applications. ResNet50 delivers competitive accuracy (92.10%) but at a higher computational cost (4.10 GFLOPs). The increased inference latency of SGAP-Net (15.54 ms) is primarily attributed to on-the-fly CLAHE preprocessing, not the model architecture itself.

H. Discussion

SGAP-Net achieves 88.15% accuracy and 97.33% AUC-ROC, driven primarily by training refinements, CLAHE, structured augmentation, label smoothing, and class weighting, rather than architectural changes. This highlights the importance of careful data preparation in medical AI, but also exposes key limitations. The model performs well on diabetic retinopathy (93.09% sensitivity) and glaucoma (91.67%), yet shows reduced sensitivity for normal cases (79.18%), leading to false positives. This issue is likely amplified by dataset imbalance: normal images originate solely from ODIR, whereas pathological samples come from multiple sources, introducing domain-specific artifacts that the model may interpret incorrectly.

The unsupervised CDR estimation pipeline, relying on Frangi filtering and Hough circle detection, is also vulnerable to anatomical variability, contributing to misclassification. Despite strong validation metrics, overfitting remains a concern because training and evaluation were conducted on internally combined public datasets without external validation. The absence of multi-center or cross-device testing further limits generalizability.

In terms of contribution, SGAP-Net employs a practical variant of existing attention mechanisms, a lightweight saliency-based enhancement module, integrated with preprocessing, augmentation, and explainability steps aligned with a clinical biomarker (retinal vasculature). While modest in technical novelty, this design improves clinical plausibility, a critical factor in medical AI.

V. CONCLUSION

SGAP-Net shows that performance and interpretability in retinal disease classification can be improved through a training pipeline that incorporates preprocessing (e.g., CLAHE), structure-preserving augmentations, and class-balanced learning, all while using a lightweight EfficientNet-B3 backbone. The model achieves 88.15% accuracy and demonstrates attention patterns that align quantitatively with retinal vasculature, as measured by ROS, RR, and MAI metrics on a consolidated test set from public datasets (ODIR, MESSIDOR, HRF).

However, several limitations must be acknowledged. First, the evaluation is confined to retrospective, publicly available data without prospective clinical validation. Second, normal-class sensitivity remains suboptimal (79.18%), leading to false positives that may limit screening utility. Third, the heuristic CDR estimation pipeline is vulnerable to anatomical variation and image quality. Fourth, inference latency (15.54 ms/image) is inflated by on-the-fly CLAHE, which is unnecessary if preprocessing is applied offline. Finally, the model has not been tested on external, multi-center data, raising concerns about generalizability across imaging devices and populations.

Future work should address these gaps through: (1) improved modeling of normal fundus appearance using uncertainty-aware or contrastive losses; (2) learned segmentation of optic disc and cup instead of rule-based methods; (3) offline preprocessing to reduce latency; (4) cross-dataset and cross-institutional validation; and (5) collaboration with clinicians to evaluate whether attention maps reflect actual diagnostic reasoning.

REFERENCES

[1] C. Zhou *et al.*, "Visual impairment and blindness caused by retinal diseases: A nationwide register-based study.," *J. Glob. Health*, vol. 13, p. 4126, Nov. 2023, doi: 10.7189/jogh.13.04126.

[2] A. Dosovitskiy *et al.*, "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE," *Int. Conf. Learn. Represent.*, pp. 1–21, 2021.

[3] Z. Shao *et al.*, "TransMIL: transformer based correlated multiple instance learning for whole slide image classification," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, in NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021.

[4] F. Shamshad *et al.*, "Transformers in medical imaging: A survey," *Med. Image Anal.*, vol. 88, no. April, p. 102802, 2023, doi: 10.1016/j.media.2023.102802.

[5] D. Luo and L. Shen, "Vessel-Net: A Vessel-aligned Ensemble Network For Retinopathy Screening From Fundus Image," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 320–324. doi: 10.1109/ICIP40778.2020.9190800.

[6] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in

Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII, Berlin, Heidelberg: Springer-Verlag, 2018, pp. 3–19. doi: 10.1007/978-3-030-01234-2_1.

[7] F. Uslu and M. Varela, "A robust quality estimation method for medical image segmentation with small datasets," *Biomed. Signal Process. Control*, vol. 95, no. PB, p. 106300, 2024, doi: 10.1016/j.bspc.2024.106300.

[8] S. Das, S. Chakraborty, M. Mishra, and S. Majumder, "Assessment of retinal blood vessel segmentation using U-Net model: A deep learning approach," *Franklin Open*, vol. 8, no. July, p. 100143, 2024, doi: 10.1016/j.fraope.2024.100143.

[9] A. Esteva *et al.*, "Deep learning-enabled medical computer vision," *npj Digit. Med.*, vol. 4, no. 1, pp. 1–9, 2021, doi: 10.1038/s41746-020-00376-2.

[10] J. Yang, C. Wang, L. Yang, Y. Jiang, and A. Cao, "Adaptive Feature Learning for Unbiased Scene Graph Generation," *Trans. Img. Proc.*, vol. 33, pp. 2252–2265, Jan. 2024, doi: 10.1109/TIP.2024.3374644.

[11] B. Zhan, E. Song, and H. Liu, "FSA-Net: Rethinking the attention mechanisms in medical image segmentation from releasing global suppressed information," *Comput. Biol. Med.*, vol. 161, no. April, p. 106932, 2023, doi: 10.1016/j.compbiomed.2023.106932.

[12] A. Wollek *et al.*, "Attention-based Saliency Maps Improve Interpretability of Pneumothorax Classification.," *Radiol. Artif. Intell.*, vol. 5, no. 2, p. e220187, Mar. 2023, doi: 10.1148/ryai.220187.

[13] A. Adadi and M. Berrada, "Explainable AI for Healthcare: From Black Box to Interpretable Models," in *Embedded Systems and Artificial Intelligence*, V. Bhateja, S. C. Satapathy, and H. Satori, Eds., Singapore: Springer Singapore, 2020, pp. 327–337.

[14] T. F. Tan *et al.*, "Explainable artificial intelligence in ophthalmology," *Curr. Opin. Ophthalmol.*, vol. 34, no. 5, 2023, [Online]. Available: https://journals.lww.com/co-ophthalmology/fulltext/2023/09000/explainable_artificial_intelligence_in.13.aspx

[15] Y. Singh *et al.*, "Beyond Post hoc Explanations: A Comprehensive Framework for Accountable AI in Medical Imaging Through Transparency, Interpretability, and Explainability," *Bioengineering*, vol. 12, no. 8, pp. 1–18, 2025, doi: 10.3390/bioengineering12080879.

[16] Y. Zhong, T. Chen, D. Zhong, and X. Liu, "Vessel Segmentation in Fundus Images with Multi-Scale Feature Extraction and Disentangled Representation," *Appl. Sci.*, vol. 14, no. 12, 2024, doi: 10.3390/app14125039.

[17] S. F. Nimmy, O. K. Hussain, R. K. Chakrabortty, and S. Saha, "Explainable Artificial Intelligence (XAI) in glaucoma assessment: Advancing the frontiers of machine learning algorithms," *Knowledge-Based Syst.*, vol. 316, no. January, 2025, doi: 10.1016/j.knosys.2025.113333.

[18] J. Guo, X. Guo, Q. Yi, H. Hao, and Y. Zhao, "Retinal vessel segmentation driven by structure prior tokens.," *Med. Phys.*, vol. 52, no. 8, p. e18018, Aug. 2025, doi: 10.1002/mp.18018.

[19] G. Saxena, D. K. Verma, A. Paraye, A. Rajan, and A. Rawat, "Improved and robust deep learning agent for preliminary detection of diabetic retinopathy using public datasets," *Intell. Med.*, vol. 3–4, no. June, p. 100022, 2020, doi: 10.1016/j.ibmed.2020.100022.

[20] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, and J. Wexler, "Interpretability Beyond Feature Attribution : Quantitative Testing with Concept Activation Vectors (TCAV)," in *International Conference on Machine Learning*, 2018.

[21] M. W. Nadeem, H. G. Goh, M. Hussain, S.-Y. Liew, I. Andonovic, and M. A. Khan, "Deep Learning for Diabetic Retinopathy Analysis: A Review, Research Challenges, and Future Directions.," *Sensors (Basel).*, vol. 22, no. 18, Sep. 2022, doi: 10.3390/s22186780.

[22] K. Jin and J. Ye, "Artificial intelligence and deep learning in ophthalmology: Current status and future perspectives," *Adv. Ophthalmol. Pract. Res.*, vol. 2, no. 3, p. 100078, 2022, doi: https://doi.org/10.1016/j.aopr.2022.100078.

[23] Q. V. Le Mingxing Tan, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *International Conference on Machine Learning*, 2019, p. 190.

[24] H. Moayed and E. G. Mansoori, "Improving Regularization in Deep Neural Networks by Co-adaptation Trace Detection," *Neural Process. Lett.*, vol. 55, no. 6, pp. 7985–7997, May 2023, doi: 10.1007/s11063-023-11293-2.

[25] M. Vadduri and P. Kuppusamy, "Enhancing Ocular Healthcare: Deep Learning-Based Multi-Class Diabetic Eye Disease Segmentation and Classification," *IEEE Access*, vol. 11, pp. 137881–137898, 2023, doi: 10.1109/ACCESS.2023.3339574.

[26] S. Vujosevic *et al.*, "Screening for diabetic retinopathy: new perspectives and challenges," *Lancet Diabetes Endocrinol.*, vol. 8, no. 4, pp. 337–347, 2020, doi: https://doi.org/10.1016/S2213-8587(19)30411-5.

[27] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in *Medical Image Computing and Computer-Assisted Intervention --- MICCAI'98*, W. M. Wells, A. Colchester, and S. Delp, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 130–137.

[28] Y.-S. Cho, H.-J. Song, J.-H. Han, and Y.-S. Kim, "Attention Mechanism-Based Glaucoma Classification Model Using Retinal Fundus Images.," *Sensors (Basel).*, vol. 24, no. 14, Jul. 2024, doi: 10.3390/s24144684.