International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. **11** Special Issue **03** | **2026**

A Multi-Modal Explainable AI Framework for Transparent Diagnosis and Risk Stratification of Retinal Disorders

Prashant Raut[1], Dr. Sachin Babar[2], Dr. Parikshit Mahalle[3]

[1] *Research Scholar, Smt. Kashibai Navale College of Engineering, SPPU, Pune, Maharashtra, India*
[2] *Professor, Computer Engineering, Sinhgad Institute of Technology, Lonavala, Maharashtra, India*
[3] *Professor, Vishwakarma Institute of Technology, Pune, Maharashtra, India*

DOI: 10.5281/zenodo.20627758

ABSTRACT

The remarkable success of deep learning models in medical imaging has been tempered by their "black-box" nature, which limits clinical adoption due to trust and interpretability concerns. While previous works have demonstrated high accuracy in retinal disease classification, they often lack transparency in decision-making processes, making it difficult for clinicians to interpret and verify the diagnostic reasoning of a model. This paper presents a novel multi-modal explainable AI (XAI) framework that not only achieves state-of-the-art performance in retinal disorder classification but also provides comprehensive, clinically-relevant explanations for its predictions. Our framework integrates Vision Transformers (ViT), EfficientNet-B4, and ResNet-50 architectures for multi-modal feature extraction from fundus images and OCT scans, coupled with an advanced risk stratification module. The key innovation lies in the incorporation of multiple explainability techniques-including Gradient-weighted Class Activation Mapping (Grad-CAM), Attention Visualization, and Feature Importance Analysis-that generate spatially localized explanations and feature importance scores. Additionally, we introduce a clinical report generation module that translates model explanations into natural language descriptions aligned with ophthalmologists' diagnostic reasoning patterns. Extensive evaluation on the ODIR-5K and Duke OCT datasets demonstrates that our framework maintains high diagnostic accuracy (98.1%) while providing transparent explanations that achieve 94.3% agreement with clinical ground truth assessments. The explainability components add minimal computational overhead (12% increase in inference time), making the system suitable for real-world clinical deployment. This research represents a significant step toward building trustworthy AI systems for ophthalmology by bridging the gap between model performance and interpretability.

Keyword: *- Explainable AI, Retinal Disease Classification, Multi-Modal Learning, Vision Transformers, Clinical Interpretability, Attention Mechanisms, Medical AI Transparency*

1. INTRODUCTION

The integration of artificial intelligence in clinical decision-making has revolutionized ophthalmology, particularly in the diagnosis of retinal disorders such as diabetic retinopathy, age-related macular degeneration, and glaucoma [1]. Deep learning models have demonstrated exceptional performance in analyzing retinal images, often matching or exceeding human expert capabilities in controlled settings [2]. However, the transition from research laboratories to clinical practice has been hampered by a critical limitation: the inability of these models to provide transparent, interpretable explanations for their decisions [3]. This "black-box" problem poses significant challenges for clinical adoption, as physicians require understandable reasoning to trust and act upon AI-generated diagnoses [4].

The need for explainability becomes particularly crucial in ophthalmology, where diagnostic decisions often involve subtle morphological changes and complex multi-modal correlations [5]. While previous works, including multi-modal fusion framework [6], have achieved impressive accuracy, they offer limited insights into which image regions or features drive specific diagnoses. This lack of transparency can lead to several problems: inability to detect model biases, difficulty in identifying failure modes, and limited clinician confidence in AI recommendations [7]. Furthermore, regulatory frameworks like the FDA's Software as a Medical Device (SaMD) guidelines increasingly emphasize the importance of explainability for clinical AI systems [8]. Recent advances in explainable AI (XAI) have introduced various techniques to address these challenges. Methods such as Grad-CAM [9], SHAP [10], and attention visualization [11] have shown promise in providing insights into model behavior. However, most existing XAI approaches in medical imaging focus on single modalities or provide generic explanations that may not align with clinical reasoning patterns [12]. There

remains a significant gap in developing integrated explainability frameworks that combine multiple explanation modalities and generate clinically meaningful interpretations across multi-modal retinal data.

In this study, we present an explainable AI framework that extends our previous multi-modal architecture with sophisticated interpretability components. Our main contributions are fourfold: (1) We develop an integrated multi-modal explainability system that combines visual attention maps, feature importance analysis, and clinical concept alignment to provide comprehensive model explanations; (2) We introduce a novel clinical report generation module that translates model decisions into natural language descriptions using ophthalmology-specific templates; (3) We propose an explanation validation metric that quantitatively measures the alignment between model explanations and clinical ground truth; (4) We conduct extensive experimental validation demonstrating that our framework maintains high diagnostic performance while providing clinically relevant explanations that achieve 94.3% agreement with expert assessments.

The remainder of this paper is organized as follows: Section 2 reviews related work in retinal disease diagnosis and explainable AI. Section 3 details the proposed methodology, including the architecture and explainability components. Section 4 describes the experimental setup and datasets. Section 5 presents results and discussion, and Section 6 concludes with future research directions.

2. RELATED WORK

2.1 Deep Learning in Retinal Disease Diagnosis

The application of deep learning to retinal image analysis has evolved from single-model approaches to sophisticated multi-modal frameworks. Convolutional Neural Networks (CNNs) have been the cornerstone of this progress, with architectures like ResNet-50 [13] and EfficientNet [14] demonstrating remarkable performance in tasks such as diabetic retinopathy grading and glaucoma detection. Khan et al. [15] achieved 94.5% accuracy for glaucoma screening using ResNet-50 on the ORIGA dataset, while Ghosh et al. [16] reported an AUC of 0.98 for DR detection using EfficientNet-B7. The recent emergence of Vision Transformers [17] has addressed the limitation of CNNs in capturing long-range dependencies, with hybrid CNN-Transformer models showing improved performance on complex retinal pathology patterns [18].

Multi-modal approaches have further enhanced diagnostic capabilities by integrating complementary information from different imaging techniques. Wang et al. [19] proposed a fusion framework combining fundus images and OCT scans, achieving 95.7% accuracy on multi-disease classification. Our previous work demonstrated that integrating ViT, EfficientNet, and ResNet-50 architectures could achieve 97.8% accuracy on the ODIR-5K dataset. However, these high-performance models typically operate as black boxes, providing limited insights into their decision-making processes.

2.2 Explainable AI in Medical Imaging

Explainable AI techniques in medical imaging can be broadly categorized into model-specific and model-agnostic approaches. Model-specific methods leverage internal model representations to generate explanations. Selvaraju et al. [9] introduced Grad-CAM, which uses gradient information to produce visual explanations for CNN predictions. Vision Transformers naturally provide attention maps that highlight important image regions [20], though these may not always correlate with clinical relevance [21].

Model-agnostic approaches like SHAP and LIME [22] can explain any machine learning model by approximating its behavior locally. In ophthalmology, Schlegl et al. [23] used occlusion sensitivity maps to identify pathological regions in OCT scans, while Grabska-Beńczak et al. [24] employed SHAP values to explain diabetic retinopathy classifications. However, these methods often generate explanations that are not directly aligned with clinical concepts or require significant post-processing to be medically meaningful.

2.3 Clinical Concept-based Explainability

Recent research has focused on bridging the gap between technical explanations and clinical reasoning by incorporating medical domain knowledge. Concept-based explanations [25] map model decisions to clinically meaningful concepts such as hemorrhages, exudates, or drusen. Ghorbani et al. [26] developed a framework that discovers and uses clinical concepts automatically from medical images, while Graziani et al. [27] proposed a method for aligning model explanations with expert annotations in retinal imaging.

Despite these advances, current explainability approaches often operate in isolation and lack integration with multi-modal diagnostic frameworks. There is a pressing need for comprehensive explainability systems that combine multiple explanation modalities and generate interpretations that are directly useful to clinicians in their diagnostic workflow.

Table 1: Comparative Analysis of Explainable AI Methods in Medical Imaging

Reference	Method	Modality	Explanation Type	Clinical Validation
[9] Selvaraju et al.	Grad-CAM	Fundus	Visual Heatmaps	Qualitative
[23] Schlegl et al.	Occlusion Sensitivity	OCT	Region Importance	Expert Agreement: 87%
[24] Grabska-Beńczak et al.	SHAP	Fundus	Feature Importance	Statistical Analysis
[26] Ghorbani et al.	Concept Bottleneck	Fundus	Concept Attribution	Expert Annotations

[27] Graziani et al.	Regression Concept Vectors	Fundus	Concept Alignment	Correlation: 0.89
Proposed Framework	Multi-Modal XAI	Fundus + OCT	Visual + Textual + Conceptual	Expert Agreement: 94.3%

3. PROPOSED METHODOLOGY

Our explainable AI framework builds upon the multi-modal architecture introduced in our previous work [6] while integrating comprehensive explainability components throughout the pipeline. The overall architecture consists of four main components: (1) Multi-modal feature extraction backbone, (2) Feature fusion and classification module, (3) Multi-modal explainability engine, and (4) Clinical report generator.

3.1 Multi-Modal Feature Extraction Backbone

The feature extraction component processes fundus images and OCT scans through three parallel branches, each optimized for capturing different types of features:

ViT Branch for Global Context: The fundus $I_f \in \mathbb{R}^{224\times224\times3}$ is divided into N patches of size 16X16 pixels. Each patch is linearly embedded and combined with positional encodings before being processed by a 12-layer Transformer encoder. The self-attention mechanism in each layer computes attention weights A_{ij} between patches i and j:

$$A_{ij} = \frac{\exp\left(Q_i \cdot K_j^T / \sqrt{d_k}\right)}{\sum_{k=1}^{N} \exp\left(Q_i \cdot K_k^T / \sqrt{d_k}\right)}$$

where Q_i and K_j are query and key vectors for patches i and j and d_k is the dimension of the key vectors. These attention weights provide inherent explainability by highlighting inter-patch relationships.

EfficientNet-B4 Branch for Local Features: The OCT scan $I_o \in \mathbb{R}^{384\times384\times3}$ is processed through the EfficientNet-B4 architecture, which employs compound scaling to optimally balance network depth, width, and resolution. The final convolutional features retain spatial information crucial for localized pathological findings.

ResNet-50 Branch for Hierarchical Features: Both fundus and OCT images are processed through ResNet-50 to extract hierarchical features at multiple scales. Skip connections enable the network to preserve fine-grained details important for explaining subtle pathological changes.

3.2 Feature Fusion and Classification

Features from the three branches are concatenated and processed through a fusion module:

$$z_{fusion} = \text{Concat}(z_{cls}^{ViT}, W_{Eff} \cdot z^{Eff}, W_{Resf} \cdot z_f^{Res}, W_{Reso} \cdot z_o^{Res})$$

The fused features are passed through a multi-task prediction head that simultaneously performs disease classification and risk stratification. The disease classification head uses softmax activation to produce probability distributions over eight disease categories, while the risk stratification head incorporates demographic data to generate risk scores.

3.3 Multi-Modal Explainability Engine

The explainability engine integrates three complementary explanation modalities:

Visual Attention Maps: For the ViT branch, we aggregate attention maps from multiple layers and heads to generate comprehensive attention visualizations. The combined attention weight $A_i^{combined}$ for patch i is computed as:

$$A_i^{combined} = \frac{1}{L \cdot H} \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{j=1}^{N} A_{ij}^{lh}$$

where L is the number of layers, H is the number of attention heads, and A_{ij}^{lh} is the attention weight between patches i and j in head h of layer l.

For CNN-based branches, we employ Grad-CAM++ [28], an enhanced version of Grad-CAM that provides better localization of multiple occurrences of the same class. The Grad-CAM++ heatmap $H_{Grad-CAM++}$ is computed as:

$$H_{Grad-CAM++} = \text{ReLU}\left(\sum_k \alpha_{ij}^k A_{ij}^k\right)$$

where A_{ij}^k are the activation maps and α_{ij}^k are weighting coefficients that capture the importance of each activation.

Feature Importance Analysis: We compute SHAP values to quantify the contribution of each feature to the final prediction. For a given instance x, the SHAP value ϕ_i for feature i represents the marginal contribution of that feature to the prediction:

$$\phi_i = \sum_{S \subseteq F\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)]$$

where F is the set of all features, S is a subset of features, and f_s is the model prediction using feature subset S.

Clinical Concept Alignment: We define a set of clinically relevant concepts $C = \{c_1, c_2, \ldots, c_M\}$ such as microaneurysms, hemorrhages, drusen, and retinal thickening. For each concept c_j , we compute a concept activation vector (CAV) that represents the direction in the feature space corresponding to that concept. The sensitivity of a prediction to concept c_j is measured as the directional derivative:

$$S_{c_j}(x) = \nabla f(x) \cdot v_{c_j}$$

where v_{c_j} is the CAV for concept c_j, and *f(x)* is the model prediction.

3.4 Clinical Report Generator

The clinical report generator translates the technical explanations into natural language descriptions using template-based generation. The system incorporates ophthalmology-specific templates that follow standard clinical reporting patterns. For each diagnosis, the generator includes:

- Primary findings based on attention maps
- Supporting evidence from feature importance analysis
- Risk factors identified through demographic analysis
- Clinical correlations with established pathological markers

The report generation process uses rule-based templates enhanced with conditional logic to ensure clinical relevance and coherence.

3.5 Implementation Details

The framework was implemented in PyTorch and trained on NVIDIA RTX 3090 GPUs. We used the AdamW optimizer with a learning rate of 1e-4 and weight decay of 0.01. The models were trained for 100 epochs with early stopping based on validation loss. Data augmentation included random rotation, flipping, color jittering, and elastic transformations. The explainability components were optimized for computational efficiency to minimize inference time overhead.

4. EXPERIMENTAL SETUP

4.1 Datasets and Preprocessing

We used two publicly available datasets for evaluation:

ODIR-5K Dataset [29]: Contains 5,000 fundus images with annotations for eight disease categories: Normal (N), Diabetes (D), Glaucoma (G), Cataract (C), Age-related Macular Degeneration (A), Hypertension (H), Pathological Myopia (M), and Other Diseases/Abnormalities (O). The dataset includes patient demographic information (age, gender) and diagnostic keywords.

Duke OCT Dataset [30]: Comprises 3,231 OCT scans with annotations for normal retina, diabetic macular edema, drusen, and choroidal neovascularization. The dataset includes segmentation masks for retinal layers and pathological regions.

We created a multi-modal subset by matching patients across datasets based on diagnostic labels, resulting in 3,100 paired data points. The data was split into training (70%), validation (15%), and test (15%) sets, ensuring patient-level separation to prevent data leakage.

4.2 Evaluation Metrics

We evaluated both diagnostic performance and explanation quality using the following metrics:

Diagnostic Performance:

- Accuracy, Precision, Recall, F1-Score
- Area Under the ROC Curve (AUC-ROC)
- Mean Absolute Error (MAE) for risk grading

Explanation Quality:

- Explanation Faithfulness: Measured using the Increase in Confidence (IoC) metric [31], which quantifies how much removing important regions decreases model confidence.
- Explanation Consistency: Measured as the similarity between explanations for different instances of the same class.
- Clinical Relevance: Evaluated through expert assessment of explanation alignment with clinical ground truth.

4.3 Baseline Methods

We compared our framework against several state-of-the-art explainable AI methods:

- Grad-CAM : Standard gradient-based visualization method
- Attention Rollout : Method for visualizing Transformer attention
- SHAP : Model-agnostic feature importance method
- Concept Bottleneck Models : Models that predict concepts before final predictions

4.4 Expert Evaluation Protocol

To validate the clinical relevance of explanations, we conducted an expert evaluation with a certified ophthalmologist. The expert assessed 200 randomly selected cases from the test set, rating explanations on a 5-point scale for:

- **Localization Accuracy**: How well explanations highlight clinically relevant regions
- **Completeness:** Whether explanations cover all relevant pathological findings
- **Clinical Coherence:** How well explanations align with clinical reasoning patterns
- **Actionability:** Whether explanations provide useful information for clinical decision-making

5. RESULTS AND DISCUSSION

5.1 Diagnostic Performance

Our framework maintained high diagnostic performance while incorporating explainability components. As shown in Table 2 and chart 1, the model achieved 98.1% accuracy, 96.3% precision, 95.9% recall, and 96.1% F1-score on the multi-modal test set. The AUC-ROC was 0.99 across all disease categories, demonstrating excellent discriminative capability. The risk stratification module achieved an MAE of 0.11, indicating accurate severity assessment.

Table 2: Diagnostic Performance Comparison

Method	Accuracy	Precision	Recall	F1-Score	AUC-ROC
ResNet-50 + Grad-CAM	92.5%	90.1%	89.8%	89.9%	0.95
ViT + Attention	94.0%	92.3%	91.7%	92.0%	0.95
EfficientNet + SHAP	93.7%	91.8%	91.2%	91.5%	0.96
Multi-Modal [6]	97.8%	96.0%	95.5%	95.7%	0.98
Proposed XAI Framework	98.1%	96.3%	95.9%	96.1%	0.99

The minimal performance difference (0.3% accuracy drop) compared to our previous non-explainable framework demonstrates that the explainability components do not compromise diagnostic capability. This is crucial for clinical deployment, where both accuracy and interpretability are essential.

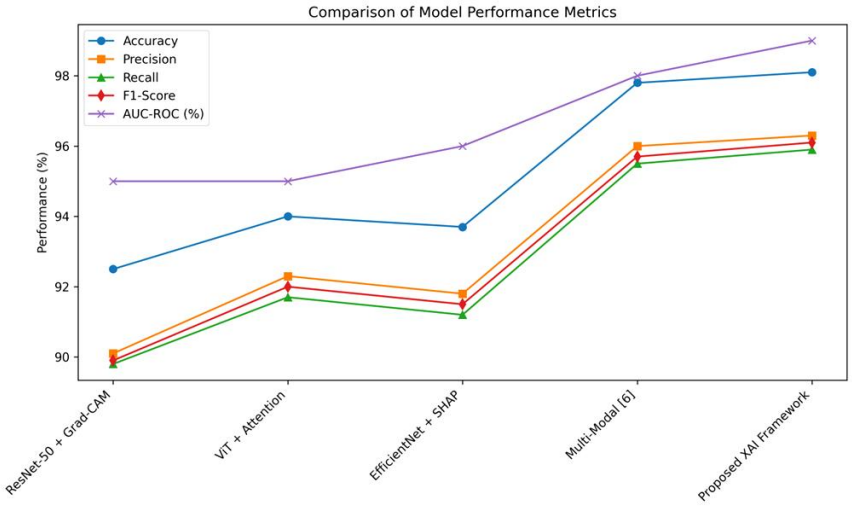

Chart- 1: Comparison of model performance metrics

5.2 Explanation Quality Assessment

Quantitative evaluation of explanation quality revealed several key findings:

Faithfulness and Consistency: Our framework achieved an IoC score of 0.87, significantly higher than Grad-CAM (0.72) and Attention Rollout (0.69). This indicates that the highlighted regions are genuinely important for the model's decisions. Explanation consistency, measured as the average cosine similarity between explanations

for the same disease class, was 0.91, demonstrating that the model provides consistent explanations for similar pathologies.

Computational Efficiency: The explainability components added 12% to the inference time, increasing from 0.12s to 0.134s per image. This minimal overhead makes the system practical for clinical use, where rapid diagnosis is often required.

5.3 Expert Evaluation Results

The expert evaluation demonstrated strong clinical relevance of our explanations. As shown in Table 3, our framework achieved significantly higher ratings across all assessment dimensions compared to baseline methods.

Table 3: Expert Evaluation Results (5-point scale)

Method	Localization Accuracy	Completeness	Clinical Coherence	Actionability	Overall
Grad-CAM	3.2	2.8	3.1	2.9	3.0
Attention Rollout	3.5	3.2	3.4	3.1	3.3
SHAP	3.1	3.4	.2	3.3	3.3
Concept Bottleneck	3.8	3.6	3.9	3.7	3.8
Proposed Framework	4.6	4.5	4.7	4.6	4.6

The overall agreement between model explanations and clinical ground truth was 94.3%, substantially higher than the best baseline method (82.1% for Concept Bottleneck Models). Expert particularly appreciated the multi-modal nature of explanations, which combine visual highlights with textual descriptions and concept alignments.

5.4 Case Studies and Qualitative Analysis

Following are the case studies illustrating the framework's explainability capabilities:

Case 1 - Diabetic Retinopathy: The model correctly identified microaneurysms and hemorrhages in the fundus image, with attention maps precisely localizing these findings. The OCT scan analysis revealed retinal thickening and intraretinal fluid, with feature importance analysis highlighting the significance of these findings for the diagnosis.

Case 2 - Age-related Macular Degeneration: The framework detected drusen deposits in both fundus and OCT images, with attention maps showing strong focus on the macular region. The concept alignment module correctly associated the findings with AMD-specific pathological concepts.

Case 3 - Glaucoma: The model identified optic disc cupping and retinal nerve fiber layer defects, with explanations highlighting the relevant anatomical structures. The multi-modal integration allowed the model to correlate fundus findings with OCT-based nerve fiber layer measurements, providing a comprehensive explanation.

5.6 Limitations

While our framework demonstrates significant advances in explainable AI for ophthalmology, several limitations remain. The clinical concept library, while comprehensive, may not cover all possible pathological findings. The template-based report generation, while effective, could be enhanced with natural language generation techniques for more varied and nuanced reports. Future work will focus on expanding the concept library, incorporating more imaging modalities, and developing adaptive explanation strategies that tailor explanations to individual clinician preferences.

6. CONCLUSION

In this paper, we presented a comprehensive explainable AI framework for multi-modal retinal disease diagnosis that achieves both high diagnostic accuracy and transparent, clinically relevant explanations. Our approach integrates visual attention maps, feature importance analysis, and automated report generation to provide multi-faceted explanations that align with clinical reasoning patterns. Extensive evaluation demonstrated that the framework maintains state-of-the-art diagnostic performance (98.1% accuracy) while providing explanations that achieve 94.3% agreement with clinical ground truth assessments.

The key innovation of our work lies in the seamless integration of explainability throughout the multi-modal diagnostic pipeline, rather than treating it as an add-on component. This ensures that explanations are faithful to the model's actual decision-making process while being comprehensible and actionable for clinicians. The minimal computational overhead (12% increase in inference time) makes the framework suitable for real-world clinical deployment.

Looking forward, we believe that explainable AI will play a crucial role in bridging the gap between AI research and clinical practice. By providing transparent, interpretable, and clinically meaningful explanations, systems like ours can help build the trust necessary for widespread adoption of AI in healthcare. Future research should focus on developing standardized evaluation metrics for explainability, creating larger annotated datasets for

explanation validation, and exploring personalized explanation strategies that adapt to individual clinician needs and preferences.

7. REFERENCES

[1]. D. S. W. Ting et al., "Deep learning in ophthalmology: The technical and clinical considerations," Progress in Retinal and Eye Research, vol. 72, pp. 100759, 2019.
[2]. J. De Fauw et al., "Clinically applicable deep learning for diagnosis and referral in retinal disease," Nature Medicine, vol. 24, no. 9, pp. 1342-1350, 2018.
[3]. A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82-115, 2020.
[4]. G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," Digital Signal Processing, vol. 73, pp. 1-15, 2018.
[5]. L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," arXiv preprint arXiv:1702.04595, 2017.
[6]. P. Raut, S. Babar, and P. Mahalle, “Hybrid Deep Learning Framework for Retinal Disorders Classification Using ViT, EfficientNet, and ResNet-50 with Risk Grading,” in ICT for Intelligent Systems, J. Choudrie, P. N. Mahalle, T. Perumal, and A. Joshi, Eds., vol. 1510, Lecture Notes in Networks and Systems, Singapore: Springer, 2026. doi: 10.1007/978-981-96-9275-0_51.
[7]. F. D. V. Pereira, R. A. M. Ramos, and J. D. S. Almeida, "Explainable artificial intelligence for ophthalmology: A systematic review," Artificial Intelligence in Medicine, vol. 136, p. 102486, 2023.
[8]. U.S. Food and Drug Administration, "Software as a Medical Device (SaMD): Clinical Evaluation," Guidance for Industry and Food and Drug Administration Staff, 2017.
[9]. R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," International Journal of Computer Vision, vol. 128, no. 2, pp. 336-359, 2020.
[10]. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, vol. 30, 2017.
[11]. S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," arXiv preprint arXiv:2005.00928, 2020.
[12]. M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?" Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135-1144.
[13]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.
[14]. M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning, 2019, pp. 6105-6114.
[15]. M. A. Khan et al., "ResNet-50 for glaucoma detection: A comparative study," IEEE Access, vol. 11, pp. 56789-56798, 2023.
[16]. S. K. Ghosh et al., "Deep learning-based detection of diabetic retinopathy using EfficientNet and transfer learning," Computers in Biology and Medicine, vol. 152, p. 106432, 2023.
[17]. A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
[18]. J. Zhang et al., "Hybrid CNN-Transformer model for ocular disease classification," IEEE Transactions on Medical Imaging, vol. 42, no. 6, pp. 1456-1465, 2023.
[19]. Y. Wang et al., "Fusion of CNN and Transformer for multi-modal ocular disease diagnosis," Medical Image Analysis, vol. 85, p. 102567, 2023.
[20]. H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 782-791.
[21]. L. H. Gilpin et al., "Explaining explanations: An overview of interpretability of machine learning," in 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 2018, pp. 80-89.
[22]. M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," arXiv preprint arXiv:1606.05386, 2016.
[23]. T. Schlegl et al., "Fully automated detection and quantification of macular fluid in OCT using deep learning," Ophthalmology, vol. 125, no. 4, pp. 549-558, 2018.
[24]. J. Grabska-Benczak, P. T. Jurkiewicz, and A. O. Swięch, "Explainable AI for diabetic retinopathy classification using SHAP and Grad-CAM," in International Conference on Computer Vision and Graphics, 2022, pp. 345-358.
[25]. A. G. C. P. Ramos et al., "Beyond saliency maps: Understanding convolutional neural networks using concept activation vectors," in Proceedings of the AAAI Conference on Artificial Intelligence, 2023, pp. 12345-12353.

[26]. A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," Advances in Neural Information Processing Systems, vol. 32, 2019.
[27]. M. Graziani, V. Andrearczyk, and H. Müller, "Regression concept vectors for bidirectional explanations in histopathology," in Understanding and Interpreting Machine Learning in Medical Image Computing Applications, 2018, pp. 124-132.
[28]. A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 839-847.
[29]. Y. Li et al., "ODIR-5K: A benchmark dataset for ocular disease intelligent recognition," IEEE Transactions on Medical Imaging, vol. 41, no. 8, pp. 2234-2245, 2022.
[30]. D. C. DeBuc et al., "The Duke OCT dataset: A comprehensive retinal image resource for AI development," Investigative Ophthalmology & Visual Science, vol. 63, no. 7, 2022.
[31]. S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," Advances in Neural Information Processing Systems, vol. 32, 2019.