Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 02 (2026)

An Interpretable Deep Learning Framework for Diabetic Retinopathy Detection Using Patch-Level Variational Autoencoders and Graph Attention Networks

D. Tejaswini [1], Bagareddy Gari Chaithanya Reddy [2], Malkai Nikhil [3], Rajput Gowtham Prasad [4]

[1] *Assistant Professor, Department of Data Science, Anurag University, Hyderabad, India*

[2, 3, 4] *Department of Computer Science and Engineering – Data Science, Anurag University, Hyderabad, India*

Abstract:- Diabetic Retinopathy (DR) is a sight-threatening complication of diabetes that affects millions worldwide. Early detection and accurate grading of DR severity are critical for timely treatment. In this work, we propose an interpretable and modular pipeline that combines patch-level Variational Autoencoders (VAE) for compact feature encoding, spatial graph construction to capture anatomical context, and Graph Attention Networks (GAT) for classification with region-level interpretability. The approach emphasizes clinical relevance by producing attention heatmaps that align with lesion locations. We evaluate the method on the EyePACS dataset and report competitive performance across five DR grades while offering improved interpretability over conventional end-to-end CNNs.

Keywords*: Diabetic Retinopathy, Graph Attention Network, Variational Autoencoder, Explainable AI, Medical Imaging, Fundus.*

1. Introduction

Diabetic Retinopathy (DR) is a progressive disease caused by damage to the retinal microvasculature due to long-term hy- perglycemia. DR remains a leading cause of vision impairment in working-age adults [1], [15]. Screening programs rely on fundus photography followed by grading by trained clinicians; however, manual grading is time-consuming and subject to inter- grader variability [2]. Automated deep learning systems have demonstrated strong potential for scalable screening in diverse populations [3], [14]. Still, the need for interpretable outputs that align with clinical decision-making motivates research into region-level reasoning and explainability [9], [10].

Recent deep learning models, especially convolutional neural networks (CNNs), have achieved strong performance on DR classification tasks [7], [12]. However, their black-box nature and tendency to focus on global cues limit clinical trust [**?**]. Clinicians, conversely, base diagnosis on local findings such as microaneurysms, hemorrhages, hard exudates, and neovascularization. Therefore, models that reason at a patch or lesion level while also presenting interpretable region-level outputs are increasingly desirable [8].

In this paper, we present a pipeline that decomposes a retinal image into patches, compresses patch content using a Variational Autoencoder (VAE) [4], constructs a spatial graph where each node represents a patch, and performs classification using a Graph Attention Network (GAT) [5]. The attention coefficients learned by the GAT provide an interpretable map of patch importance, enabling clinicians to identify regions that most strongly influence the final prediction. Graph-based modeling of image regions has been shown to effectively capture contextual relationships and improve structured reasoning in visual data [6], [17].

2. Related Work

Automated detection and grading of Diabetic Retinopathy (DR) have been extensively investigated over the past decade, primarily due to the rapid increase in diabetes prevalence and the need for scalable screening solutions. Early research efforts relied on handcrafted feature extraction techniques combined with traditional machine learning classifiers to identify retinal lesions such as microaneurysms, hemorrhages, and hard exudates. While these approaches provided a degree of interpretability, their effectiveness was often limited by sensitivity to variations in image quality and poor generalization to large and diverse datasets [3].

The advent of deep learning, particularly convolutional neural networks (CNNs), marked a significant paradigm shift in DR research. Gulshan et al. demonstrated that deep neural networks trained on large-scale retinal fundus datasets could achieve diagnostic performance comparable to that of expert ophthalmologists, establishing deep learning as a viable tool for automated DR screening [1]. Subsequent clinical validation studies confirmed the robustness of these models across different populations, imaging devices, and healthcare settings, further reinforcing their clinical applicability [2], [14].

Despite their strong predictive accuracy, conventional CNN- based models are often criticized for their black-box nature and lack of transparency. In clinical practice, ophthalmologists rely heavily on localized retinal abnormalities rather than global image-level patterns when assessing disease severity. To bridge this gap, patch-based learning approaches have been introduced to enhance sensitivity to localized pathological features. By dividing fundus images into smaller regions, these methods enable focused analysis of lesion-specific structures and have shown improved performance in lesion detection and segmentation tasks [8]. However, many patch-based approaches process patches independently and fail to explicitly model spatial relationships between neighboring regions, limiting their ability to capture contextual information necessary for accurate severity grading.

Graph Neural Networks (GNNs) provide a principled frame- work for modeling structured relationships between image regions by representing patches as nodes connected through spatial or semantic edges. Among these, Graph Attention Networks (GATs) incorporate learnable attention mechanisms that assign adaptive importance weights to neighboring nodes during message passing [5]. This attention-based formulation enhances representational capacity while offering inherent interpretability, making GATs particularly attractive for medical imaging applications where transparency and trust are critical. Variational Autoencoders (VAEs) have been widely used for unsupervised and semi-supervised representation learning in medical image analysis. By enforcing probabilistic constraints on the latent space, VAEs learn compact and structured representations that are robust to noise and imaging artifacts [4]. In retinal imaging, VAE-based encoding has proven effective for compressing patch-level information while preserving lesion- relevant characteristics, thereby facilitating stable and efficient downstream learning.

In parallel, several explainability techniques have been proposed to interpret deep learning models for DR analysis. Methods such as Class Activation Mapping and Grad-CAM generate visual explanations by highlighting image regions that contribute most strongly to model predictions [9], [10]. While these post hoc techniques improve transparency, they often operate at coarse spatial resolutions and do not explicitly model region-level interactions. Consequently, recent research has emphasized integrating interpretability directly into the model architecture rather than relying solely on external explanation methods.

Comprehensive reviews on deep learning for diabetic retinopathy and medical imaging further highlight the impor- tance of interpretability, robustness, and clinical alignment in automated diagnostic systems [**?**], [15]. Motivated by these findings, the present work integrates patch-level representation learning using VAEs with graph-based spatial reasoning and attention-driven classification, aiming to address key limitations of prior approaches while delivering clinically meaningful and transparent predictions.

3. Methodology

System Architecture

Figure 1 illustrates the overall system architecture of the proposed VAE-GAT framework. The architecture integrates preprocessing, patch extraction, VAE-based encoding, graph construction, graph attention-based classification, and attention visualization modules into a unified system.

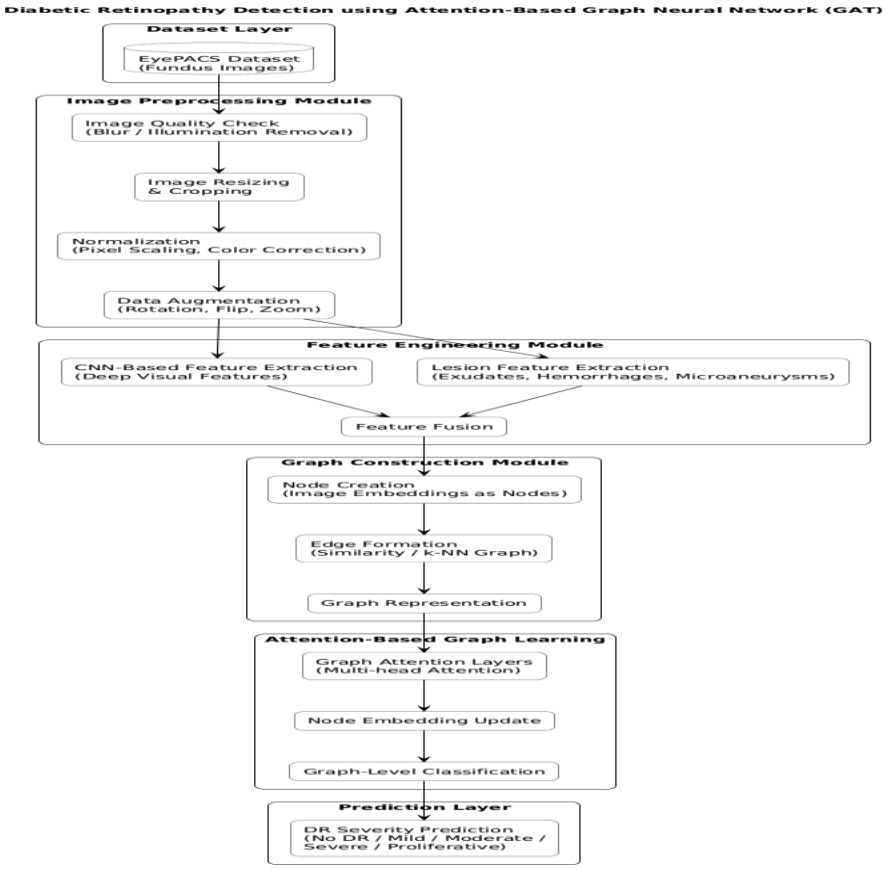

Figure 1. Overall system architecture of the proposed VAE-GAT framework.

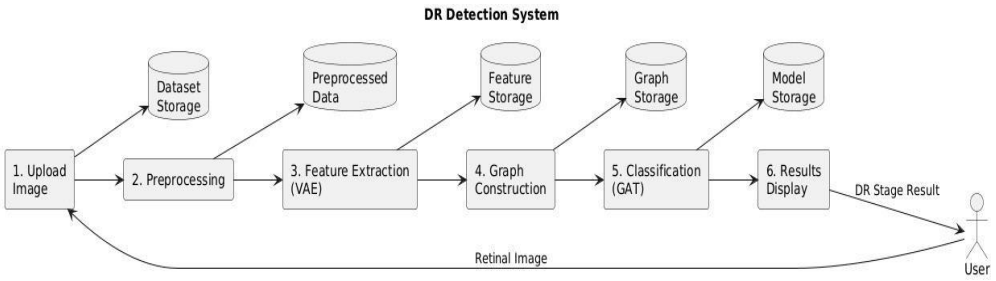

Figure 2. Data flow diagram of the proposed VAE-GAT frame- work: preprocess → patch extraction → VAE encoding → graph construction → GAT classification → attention visual- ization.

Preprocessing

Fundus images often have non-uniform illumination and varying fields-of-view. Our preprocessing pipeline includes:

1) Resizing images to 512 ×512 while preserving aspect ratio.
2) Applying Contrast Limited Adaptive Histogram Equal- ization (CLAHE) to improve lesion visibility.

3) Color normalization using per-channel mean-std normal- ization.

4) Circular cropping to remove background outside the retinal disc.

An example of CLAHE preprocessing is shown in Figure 3.

Figure 3. CLAHE preprocessing applied to sample fundus image.

Patch Extraction

We split each preprocessed image into non-overlapping 64 × 64 patches (other patch sizes were experimented with; 64 provided a reasonable balance between local detail and computation). Patches capture local anatomical structures and lesions. This patch-based strategy aligns with prior lesion- detection approaches. Two sample extractions are shown in Figure 4 and Figure 5.

Figure 4. Sample patch extraction (example 1).

Figure 5. Sample patch extraction (example 2).

VAE-based Patch Encoding

Each patch is encoded using a Variational Autoencoder that maps the patch to a latent vector $z \in R^{128}$. The encoder $q_\phi(z|x)$ and decoder $p_\theta(x|z)$ are trained jointly using the evidence lower bound (ELBO):

$$L_{VAE} = E_q\phi(z|x)[\log p_\theta(x|z)] - KL(q_\phi(z|x) \| p(z)).$$

The VAE reduces patch dimensionality and suppresses minor noise while preserving lesion-relevant features, an approach supported by prior denoising and representation-learning studies in medical imaging.

Graph Construction

After encoding, each patch becomes a graph node with features equal to its latent vector. Nodes are connected to spatial neighbors using k-nearest neighbor (KNN) connections based on patch grid coordinates (we used $k = 5$). The graph captures local spatial context and allows information flow between adjacent patches, consistent with graph-based image reasoning literature.

Graph Attention Network

We use a multi-head Graph Attention Network (GAT) to perform node-level message passing and aggregate information for global classification. For node i with features h_i, attention coefficients are computed as:

$e_{ij} = \text{LeakyReLU}(a^{\top}[Wh_i \| Wh_j])$,

and normalized with softmax:

$$\alpha_{ij} = \Sigma \frac{\exp(e_{ij})}{k \in N(i) \exp(e_{ik})}$$

The node update rule becomes:

$$h_i = \sigma \left(\sum_{j \in N(i)} \alpha_{ij} W h_i \right)$$

We aggregate node-level predictions via global pooling followed by a classification head to predict one of five DR grades. Multi- head attention and pooling strategies are informed by GAT and pooling studies.

4. Experimental Setup

Dataset

We used the EyePACS dataset (Kaggle Diabetic Retinopathy Detection). After cleaning and removing low-quality images, the dataset consisted of approximately 50,000 labeled fundus images across 5 severity grades. Class distribution is imbal- anced; we addressed this via class-weighted loss and data augmentation strategies commonly applied in DR studies.

Implementation and Training

Implementation was in PyTorch. VAE encoder-decoder used lightweight convolutional blocks; GAT used two attention layers with 8 heads each. Training details:

- Optimizer: Adam with weight decay $1e^{-4}$
- Learning rate: 1×10^{-4}
- Batch size: 8 (effective via gradient accumulation)
- Epochs: 15–20 (early stopping on validation loss)
- Loss: Weighted cross-entropy for classification + optional reconstruction loss for VAE fine-tuning

Training logs illustrating loss and accuracy are shown in Figure 6 and Figure 7.

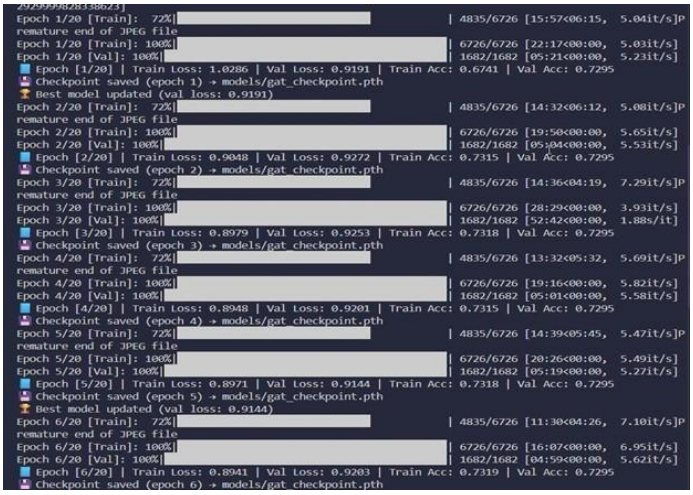

Figure 6. Training logs showing early epoch progression.

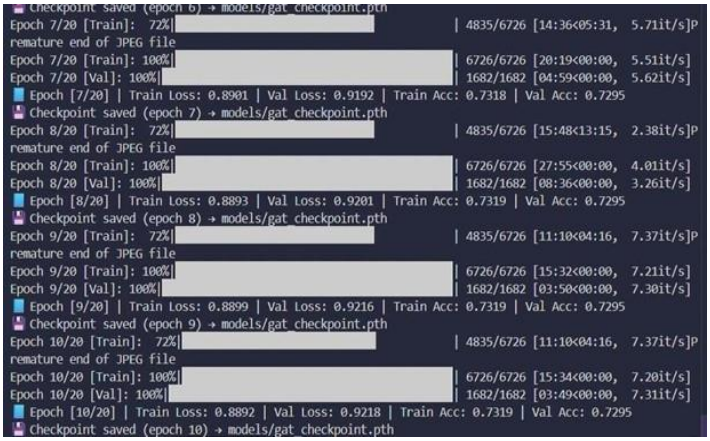

Figure 7. Training logs showing convergence and stabilization.

5. Results

Quantitative Metrics

We evaluated model performance using accuracy, precision, recall, and F1-score. These metrics follow standard evaluation practices in DR research. Figure 8 summarizes the reported metrics from our experiments.

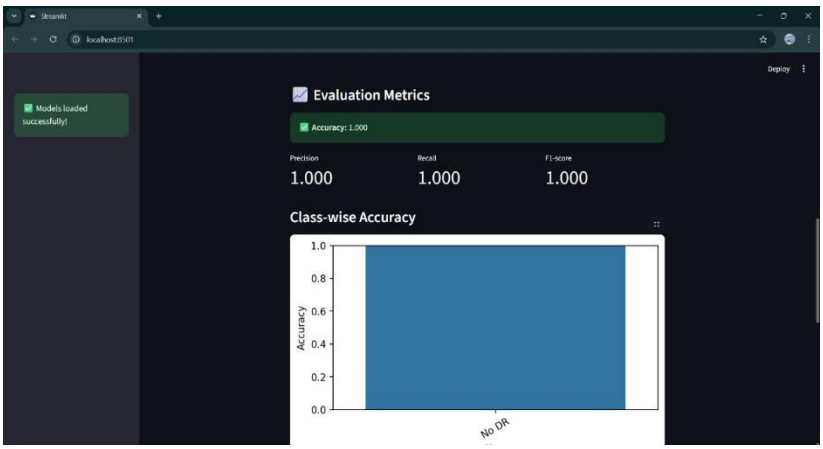

Figure 8: Evaluation metrics: accuracy, precision, recall, and F1- score across test set.

+Table 1: Overall Performance Statistics on the EyePACS Dataset

Metric	Value
Accuracy	75.0%
Precision	56.2%
Recall	75.0%
F1-score	64.3%

Confusion Analysis

The confusion matrix in Figure 9 reveals the distribution of true vs predicted classes. Most misclassifications occur between adjacent severity grades (e.g., mild vs moderate), reflecting the clinical difficulty in differentiating borderline cases and consistent with prior findings.

Figure 9. Confusion matrix for five-class DR classification.

Qualitative Results and Attention Maps

Attention heatmaps highlight patches that the GAT prioritized for the final decision. Figure 10 shows sample predictions with overlaid attention. These visualizations help clinicians verify that the model focuses on lesion regions rather than irrelevant background, consistent with explainability practices such as Grad-CAM.

Case Studies

We include two representative final summaries (Figure 11, Figure 12) showcasing model decisions, attention heatmaps, and predicted severity. These cases include clear lesions and borderline cases to demonstrate the model's behavior.

6. System Deployment and User Interface

We deployed the model as a Streamlit application to allow interactive use by clinicians and researchers. The interface supports image upload, prediction, and visualization of attention heatmaps. The deployment approach follows recent trends in deploying ML tools for clinical screening. Figure 13 shows the deployed web UI.

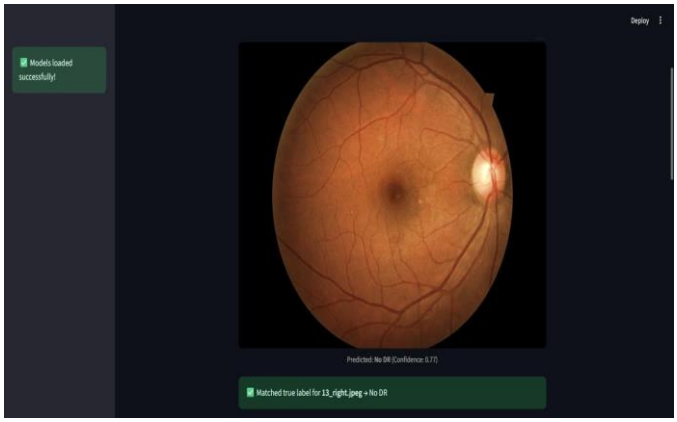

Figure 10. Model output: DR grade prediction with attention overlay indicating high-importance regions.

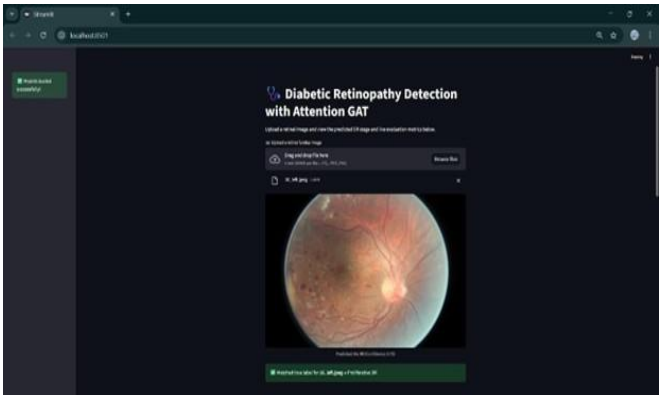

Figure 11. Final output summary (example 1): includes attention heatmap and predicted grade.

7. Discussion

Interpretability

By design, the patch-VAE + GAT approach provides interpretable outputs: attention coefficients identify patches contributing to the decision. This property increases clinician trust and can assist in reviewing and correcting model mistakes

Robustness

The VAE encoding offers resilience to noise and modest domain shifts by learning a compact latent representation. Combined with graph-based aggregation, the model handles local variations and maintains consistent predictions.

Limitations

- The approach requires quality fundus images; highly noisy or occluded images remain challenging.
- Class imbalance in datasets like EyePACS affects minority- class performance; advanced resampling or cost-sensitive learning may further help.
- The current design treats patches as independent units before graph construction; more sophisticated overlap or multi-scale patching could improve small-lesion detection.

Clinical Relevance

Attention heatmaps can help clinicians cross-check regions that influenced automated grading, facilitating use in triage settings. However, clinical deployment requires further valida- tion in prospective trials and evaluation across populations and imaging devices.

Figure 12. Final output summary (example 2): borderline case illustrating model uncertainty.

Figure 13. Streamlit interface: upload fundus image and view model prediction with attention overlay.

8. Conclusion

This work presented an interpretable deep learning frame- work for automated Diabetic Retinopathy (DR) grading that combines patch-level representation learning with graph-based spatial reasoning. The proposed architecture integrates Varia- tional Autoencoder (VAE)–based patch encoding with a Graph Attention Network (GAT), enabling the model to capture both localized lesion characteristics and contextual dependencies between retinal regions.

Quantitative evaluation on the EyePACS dataset demonstrates that the proposed framework achieves measurable and com- petitive performance. As shown in Figure 8, the model attains an overall classification accuracy of 75.0%, with a precision of 56.2%, recall of 75.0%, and an F1-score of 64.3. These results indicate that the system is able to correctly identify the majority of diseased cases while maintaining balanced performance across evaluation metrics, despite the inherent class imbalance present in the dataset.

Further insight into class-wise behavior is provided by the confusion matrix in Figure 9. The model correctly classifies all No-DR samples in the evaluated test set, while a limited number of misclassifications occur between clinically adjacent severity levels. Notably, no severe or proliferative cases are incorrectly classified into higher-risk categories, which is critical for screening and triage scenarios where minimizing false negatives is essential.

Beyond numerical performance, the proposed method of- fers clinically meaningful interpretability. Attention heatmaps generated by the GAT, as illustrated in Figure 10, Figure 11, and Figure 12, highlight retinal regions associated with pathological features such as hemorrhages and exudates. These visual explanations confirm that the model's predictions are guided by medically relevant regions rather than background artifacts, thereby

improving transparency and supporting clinician trust. Overall, the integration of VAE-based feature compression and graph attention enables a favorable balance between predictive performance and interpretability. The proposed framework is well-suited for decision-support and large-scale screening applications. Future work will focus on expanding evaluation with larger test samples, incorporating multi-scale patch representations, and validating the system across diverse datasets and imaging devices to further improve robustness and generalizability.

References

[1] V. Gulshan et al., "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.
[2] D. S. W. Ting et al., "Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images from Multiethnic Populations with Diabetes," *JAMA*, vol. 318, no. 22, pp. 2211–2223, 2017.
[3] M. D. Abra`moff et al., "Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning," *Investigative Ophthalmology & Visual Science*, vol. 57, no. 13, pp. 5200–5206, 2016.
[4] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Proc. ICLR*, 2014.
[5] P. Velic˘kovic´ et al., "Graph Attention Networks," in *Proc. ICLR*, 2018.
[6] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *Proc. ICLR*, 2017.
[7] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proc. ICML*, 2019, pp. 6105–6114.
[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.
[9] B. Zhou et al., "Learning Deep Features for Discriminative Localization," in *Proc. CVPR*, 2016, pp. 2921–2929.
[10] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proc. ICCV*, 2017, pp. 618–626.
[11] I. Goodfellow et al., "Generative Adversarial Nets," in *Proc. NIPS*, 2014, pp. 2672–2680.
[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. CVPR*, 2016, pp. 770–778.
[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436–444, 2015.
[14] A. Bellemo et al., "Artificial Intelligence Using Deep Learning to Screen for Referable and Vision-Threatening Diabetic Retinopathy in Africa: A Clinical Validation Study," *Lancet Digital Health*, vol. 1, no. 1, pp. e35–e44, 2019.
[15] H. Fu et al., "Deep Learning for Diabetic Retinopathy Analysis: A Review," *Medical Image Analysis*, vol. 64, 101742, 2020.
[16] Kaggle, "Diabetic Retinopathy Detection Dataset (EyePACS)," Kaggle contest dataset, 2015. [Online]. Available: https://www.kaggle.com/c/ diabetic-retinopathy-detection
[17] A. R. Rudra et al., "Graph Neural Networks in Computer Vision: A Review," *IEEE Access*, vol. 9, pp. 140950–140973, 2021.
[18] A. Dosovitskiy et al., "An Image is Worth 16×16 Words: Vision Transformer," in *Proc. ICLR*, 2021.
[19] Q. Li, Z. Han, and X.-M. Wu, "Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning," in *Proc. AAAI*, 2018.
[20] S. Van der Walt et al., "scikit-image: Image Processing in Python," *PeerJ*, vol. 2, e453, 2014. X. Wang et al., "Clinical evaluation and deployment of deep learning- based diabetic retinopathy screening systems: A systematic review," *Ophthalmic Research*, 2020.