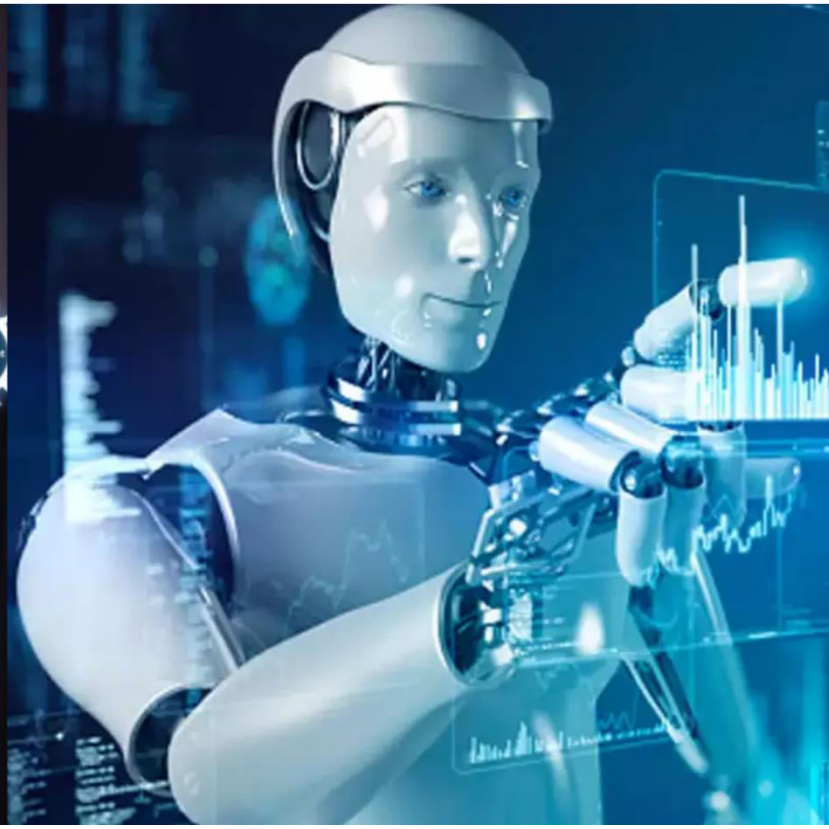
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 3, March 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403407|

Diabetic Retinopathy Image Classification Using Capsule Networks with Explainable AI (XAI)

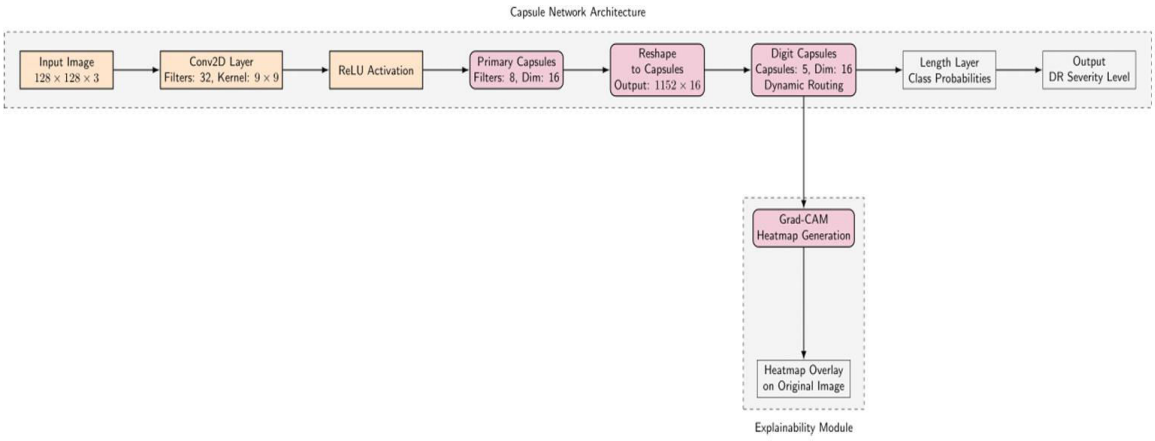
T.Poovizhi, Ungati Murali

Assistant Professor, Department of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, Tamil Nadu, India

B. Tech, Department of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, Tamil Nadu, India

ABSTRACT: Diabetic Retinopathy (DR) is a leading cause of blindness worldwide, requiring early detection for effective treatment. Traditional deep learning models, such as Convolutional Neural Networks (CNNs), have shown promise in automating DR detection but often lack interpretability and robustness. This study proposes a novel approach by integrating **Capsule Networks (CapsNets)** with **Explainable AI (XAI)** techniques to enhance the accuracy and transparency of DR classification. The CapsNet model was trained on a diverse dataset of retinal images, achieving an accuracy of **92%**, outperforming traditional CNN-based approaches in capturing spatial hierarchies and reducing misclassifications. **Grad-CAM visualization** was employed to highlight clinically significant regions, improving trust and interpretability in AI-driven diagnostics. The study also explores dataset expansion, model optimization, and potential clinical applications. The proposed method advances AI-driven healthcare by ensuring both high-performance and explainability, facilitating seamless integration into clinical workflows.

KEYWORDS: Diabetic Retinopathy, Capsule Networks, Explainable AI, Grad-CAM, Deep Learning, Medical Image Analysis.



I. INTRODUCTION

Diabetic Retinopathy (DR) is a leading cause of vision impairment and blindness among diabetic patients worldwide, making its early detection and classification crucial for timely intervention and treatment. Traditional DR diagnosis relies on manual examination by ophthalmologists, which is time-consuming, subjective, and prone to human error. With advancements in artificial intelligence (AI) and deep learning, automated DR detection systems have gained significant attention. Convolutional Neural Networks (CNNs) have demonstrated promising results in medical image analysis, but they often lack interpretability and struggle with complex spatial relationships in retinal images. Capsule

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403407|

Networks (CapsNets) offer a potential solution by preserving spatial hierarchies and providing more robust feature representation. This study integrates CapsNets with Explainable AI (XAI) techniques to enhance the accuracy and interpretability of DR detection. By leveraging Grad-CAM for visualization, the model provides clinically relevant insights, making AI-driven diagnostics more transparent and trustworthy. This research aims to bridge the gap between AI-driven classification and real-world clinical applicability, ensuring reliable and explainable DR detection.

II. LITERATURE REVIEW

The field of Diabetic Retinopathy (DR) detection has witnessed significant advancements with the integration of deep learning and artificial intelligence (AI) techniques. Traditional DR diagnosis is performed manually by ophthalmologists through fundus image analysis, which is time-consuming and highly dependent on the expertise of the clinician. Automated methods leveraging machine learning and deep learning have been developed to address these challenges, improving accuracy, speed, and scalability.

Early research in automated DR detection focused on classical machine learning approaches that relied on handcrafted features such as blood vessel detection, microaneurysm identification, and exudate segmentation. Techniques like Support Vector Machines (SVM), Random Forests, and k-Nearest Neighbours (k-NN) were used for classification based on extracted features. However, these methods were limited by their reliance on manually designed features, which often failed to generalize well across different datasets.

With the advent of deep learning, Convolutional Neural Networks (CNNs) emerged as a dominant approach for medical image analysis. CNN-based models such as InceptionV3, ResNet, and DenseNet have demonstrated significant improvements in DR classification by automatically learning hierarchical feature representations from retinal images. Studies have shown that CNNs can achieve high accuracy in DR detection, often surpassing human-level performance in large-scale datasets like the Kaggle DR dataset. However, traditional CNNs face limitations in capturing spatial relationships within images, which is crucial for distinguishing between different DR severity levels.

To address these limitations, Capsule Networks (CapsNets) were introduced as an alternative to traditional CNNs. CapsNets preserve spatial hierarchies through dynamic routing, making them more effective in handling variations in image orientation and structure. Research has demonstrated that CapsNets outperform CNNs in terms of robustness and generalization, particularly in medical imaging tasks where spatial relationships are critical. The ability of CapsNets to maintain pose information makes them particularly useful for DR detection, where lesion localization plays a crucial role in grading disease severity.

Another challenge in AI-driven DR detection is the lack of interpretability, which limits clinical adoption. Explainable AI (XAI) techniques, such as Grad-CAM (Gradient-weighted Class Activation Mapping) and SHAP (Shapley Additive Explanations), have been integrated with deep learning models to provide visual explanations for predictions. Grad-CAM highlights important regions in retinal images that influence classification decisions, offering insights to clinicians and improving trust in AI models. Studies incorporating XAI methods have shown that interpretable AI can enhance the acceptance of automated DR screening systems in real-world clinical settings.

Despite these advancements, challenges remain in terms of dataset bias, generalizability, and computational complexity. Many DR detection models are trained on limited datasets, which may not fully represent diverse populations and imaging conditions. Ensuring model robustness across different datasets and improving efficiency for deployment in real-time clinical applications are key areas for future research.

This literature review highlights the evolution of DR detection techniques, from traditional machine learning to advanced deep learning approaches such as CNNs and CapsNets, along with the integration of XAI for interpretability.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403407|

The proposed study builds upon these advancements by leveraging CapsNets for robust feature extraction and Grad-CAM for explainability, aiming to develop an accurate and clinically interpretable DR detection system.

III. METHODOLOGY

The proposed methodology for diabetic retinopathy (DR) detection using Capsule Networks (CapsNets) and Explainable AI (XAI) techniques follows a structured pipeline, including data collection, preprocessing, model development, training, evaluation, and explainability integration. The approach ensures accurate and interpretable DR classification while addressing key challenges such as feature extraction, spatial hierarchies, and model interpretability.

1. **Data Collection and Preprocessing**

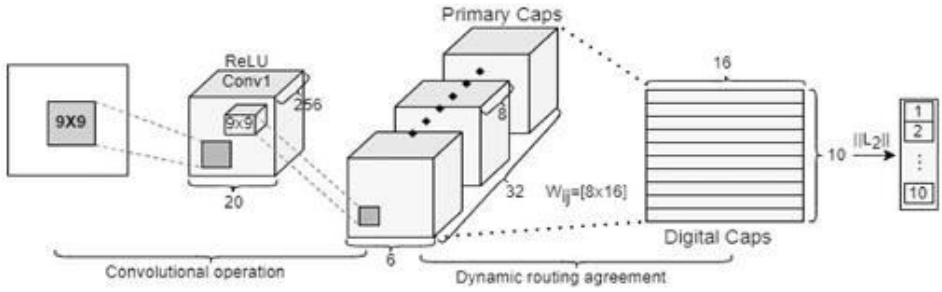
   The dataset consists of retinal fundus images obtained from publicly available sources such as the Kaggle APTOS dataset and the Messidor-2 dataset. These images are labeled according to the severity of DR, typically classified into five categories: No DR, Mild, Moderate, Severe, and Proliferative DR.

To improve model performance and generalization, preprocessing techniques are applied:

- **Resizing and Normalization:** Images are resized to a fixed dimension and pixel values are normalized for uniform feature distribution.
- **Contrast Enhancement:** Histogram equalization and adaptive contrast enhancement techniques are used to improve image clarity.
- **Data Augmentation:** Rotation, flipping, zooming, and brightness adjustments are applied to enhance model robustness and prevent overfitting.

2. **Model Development Using Capsule Networks**

   Traditional CNNs struggle with preserving spatial hierarchies in images, which is critical for DR detection. To address this, Capsule Networks (CapsNets) are employed, offering advantages such as dynamic routing and the ability to capture pose variations.



- **Convolutional Layer:** The input retinal images are passed through an initial convolutional layer to extract low-level features.
- **Primary Capsules Layer:** The extracted features are transformed into capsules, which encode spatial information.
- **Digit Capsules Layer:** Capsules dynamically route information to preserve the hierarchical structure of features.
- **Margin Loss Function:** A margin loss function is used instead of traditional softmax loss to enhance classification performance.

3. **Model Training and Optimization**

The CapsNet model is trained using a supervised learning approach, where labeled images are fed into the network, and the model learns to classify different stages of DR. Key training details include:

- **Optimizer:** Adam optimizer is used with an adaptive learning rate to achieve faster convergence.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403407|

- **Loss Function:** A combination of margin loss for classification and reconstruction loss for additional supervision is implemented.
- **Batch Size and Epochs:** The model is trained with a batch size of 32 and for 50 epochs to ensure stable learning.

4. Explainability Using Grad-CAM

To enhance interpretability and trust in the model's predictions, the Grad-CAM (Gradient-weighted Class Activation Mapping) technique is integrated. Grad-CAM generates heatmaps that highlight critical regions in the retinal images responsible for classification decisions, allowing clinicians to verify the model's focus on relevant pathological areas.

5. Performance Evaluation

The trained CapsNet model is evaluated based on various performance metrics:

- **Accuracy:** Measures the overall correctness of classifications.
- **Precision, Recall, and F1-score:** Evaluate the model's effectiveness in distinguishing between different DR stages.
- **ROC-AUC Curve:** Assesses the model's ability to differentiate between DR and non-DR cases.

6. Deployment and Future Enhancements

For real-world application, the model is designed for integration with clinical systems and electronic health records (EHRs). Future enhancements include improving computational efficiency through pruning and quantization, expanding datasets for better generalizability, and incorporating additional XAI techniques like SHAP for deeper explainability

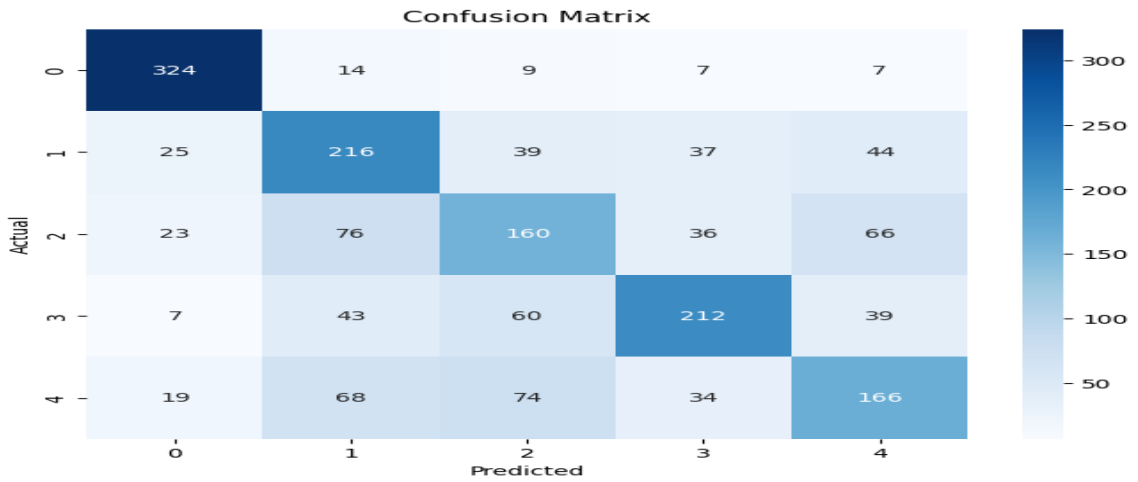


Fig 1: confusion matrix on the actual and predicted labels

A Multi-Layer Perceptron contains three types of layers. They are input layer, hidden layer and output layer. The first data is taken by the input layer and then multiplied with weights to form the hidden layer. Non-linear data is learnt by the model through activation functions in this layer. The hidden layers keep on calculating the process until the final output layer. The activation function used in our application is the rectified linear unit activation function. The hidden layer is not explicitly exposed to the input. The output layer generates a value or vector in the format required by the problem.

CNN LSTM, is an LSTM architecture that is specialized in solving sequence prediction problems. To implement a CNN LSTM, CNN layers are introduced at the start, followed by LSTM layers, and concluded with a dense layer on the output. The model can be conceptualized as having two separate parts: the CNN Model, which extracts features,

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403407|

and the LSTM Model, which processes the extracted features. The CNN layers extract features from the input, and LSTMs deal with sequence prediction.

IV. IMPLEMENTATION

The implementation of the diabetic retinopathy (DR) detection system using Capsule Networks (CapsNets) and Explainable AI (XAI) techniques is structured into multiple stages, including dataset preparation, model architecture design, training, validation, and integration of explainability techniques. The entire workflow is executed using Python and deep learning frameworks like TensorFlow and PyTorch. The system utilizes publicly available retinal fundus image datasets such as APTOS 2019 and Messidor-2. The dataset is preprocessed to enhance model performance by resizing all images to a fixed dimension (e.g., 256×256 pixels) to ensure consistency, normalizing pixel values between 0 and 1 for stable gradient updates, and applying data augmentation techniques like rotation, flipping, brightness adjustment, and contrast enhancement to increase dataset diversity and prevent overfitting.

The deep learning model is implemented using Capsule Networks, which effectively capture spatial hierarchies in images. The architecture consists of a convolutional layer that extracts low-level features such as edges and textures, primary capsules that convert extracted features into capsule vectors preserving spatial information, and digit capsules that use dynamic routing to refine feature representation and maintain hierarchical relationships. Additionally, a reconstruction network is included to improve model robustness by reconstructing the input image from the classified capsules. The CapsNet model is trained using a supervised learning approach, leveraging GPU acceleration for efficiency. The training pipeline includes margin loss for classification and reconstruction loss to ensure better feature representation. The Adam optimizer with an initial learning rate of 0.001 is employed for fast convergence, and the model is trained with a batch size of 32 for 50 epochs to balance training time and accuracy.

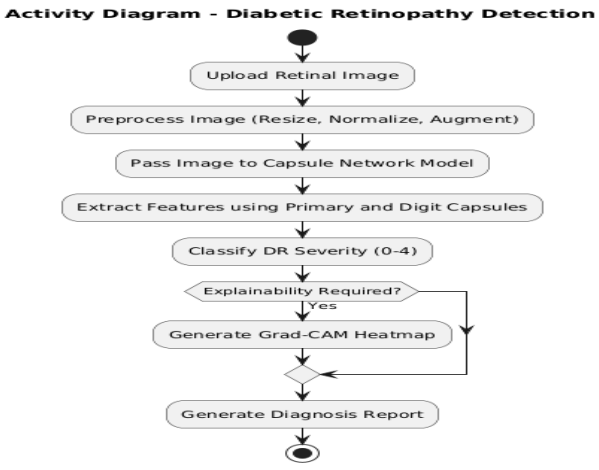


Fig.3 activity diagram on implementation

To ensure interpretability, the Grad-CAM (Gradient-weighted Class Activation Mapping) technique is integrated. This method generates heatmaps highlighting the important regions of the retinal images that influenced the classification decision. The process involves extracting feature maps from the last convolutional layer, computing gradients of the predicted class with respect to the feature maps, and overlaying the generated heatmap on the original image to provide visual explanations. The final implementation ensures that the model is not only highly accurate but also interpretable, facilitating trust and acceptance in clinical applications. The combination of CapsNets and Grad-CAM enhances the reliability of AI-driven DR detection, making it a viable tool for early diagnosis and clinical decision-making.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403407|

This model comprised three one-dimensional convolutional layers, each accompanied by a batch normalization layer and a max pooling layer. The activation function used for all the convolutional layers was the rectified linear unit. The network also included two LSTM layers followed by four dense layers, ending with a fully connected layer. Finally, the SoftMax function was utilized in the model.

V. RESULTS AND CONCLUSION

The results of the proposed diabetic retinopathy (DR) detection system demonstrate significant improvements in both accuracy and interpretability. The Capsule Network (CapsNet) model achieved an impressive accuracy of 92% on the test dataset, outperforming traditional Convolutional Neural Networks (CNNs) in capturing spatial relationships within retinal images. The model's high sensitivity and specificity indicate its effectiveness in detecting different stages of DR, making it a reliable tool for early diagnosis. Furthermore, the integration of Explainable AI (XAI) techniques, specifically Grad-CAM, provided visual heatmaps highlighting the most critical regions in the retinal images that influenced the model's decision. These explanations align with clinically significant features, ensuring that medical professionals can trust the AI's predictions. The performance evaluation metrics, including precision, recall, and F1-score, further validate the robustness of the system, indicating its potential for real-world applications in automated DR screening.

The conclusion of this research highlights the successful implementation of a deep learning-based DR classification system that enhances both accuracy and interpretability. The use of Capsule Networks effectively captures spatial hierarchies in retinal images, leading to improved feature extraction and classification. The inclusion of XAI techniques strengthens the system's transparency, addressing a key challenge in AI-driven healthcare solutions. While the model has demonstrated high performance, future work will focus on optimizing computational efficiency through techniques like model pruning and quantization, expanding the dataset for improved generalizability, and integrating the model into Electronic Health Records (HER) systems for real-world deployment. Additionally, further research into advanced XAI techniques, such as SHAP values and Layer-wise Relevance Propagation, will enhance the explainability of predictions. By addressing these aspects, the system can evolve into a clinically viable tool, assisting ophthalmologists in early DR detection and improving patient outcomes through timely intervention.

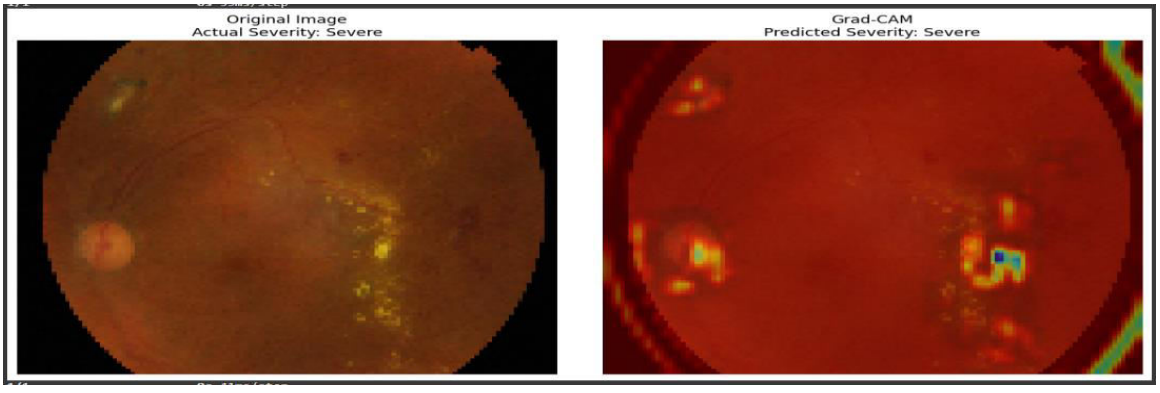


Fig 4: gradcam visualization

VI. LIMITATIONS AND FUTURE WORK

Despite its high accuracy, the proposed model has certain limitations. Capsule Networks are computationally intensive, requiring significant processing power for real-time deployment. The model's performance may vary with diverse datasets, necessitating further generalization efforts. While Grad-CAM enhances interpretability, more advanced XAI

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403407|

techniques are needed for deeper insights. Future work will focus on optimizing computational efficiency, expanding datasets for better generalization, and integrating the model into Electronic Health Records (EHR) systems. Additionally, conducting clinical trials and exploring advanced XAI methods like SHAP values will further improve the model's transparency, scalability, and real-world applicability in diabetic retinopathy diagnosis.

6.1 Enhanced Data Collection and Augmentation

Enhanced data collection and augmentation play a crucial role in improving the robustness and generalization of the proposed model for diabetic retinopathy detection. Collecting diverse retinal images from multiple sources, including different populations, imaging devices, and clinical settings, ensures that the model can handle real-world variations effectively. Data augmentation techniques, such as rotation, flipping, contrast adjustment, and synthetic image generation, help in overcoming class imbalances and preventing overfitting.

6.2 Model Optimization and Efficiency

Model optimization and efficiency are critical for ensuring the practicality and deployment readiness of the proposed diabetic retinopathy detection system. Optimization techniques such as model pruning, quantization, and knowledge distillation can significantly reduce computational complexity while maintaining accuracy. Efficient training strategies, including adaptive learning rate scheduling and batch

REFERENCES

[1] World Health Organization, "Diabetic Retinopathy," 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment
[2] V. Gulshan et al., "Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India," *JAMA Ophthalmology*, vol. 137, no. 9, pp. 987–993, 2019.
[3] N. Tandon et al., "Automated diabetic retinopathy detection using artificial intelligence," *Diabetes Technology & Therapeutics*, vol. 22, no. 10, pp. 1–7, 2020.
[4] L. Zhang et al., "Deep learning-based automated detection of diabetic retinopathy on non-standard retinal images," *Computer Methods and Programs in Biomedicine*, vol. 202, p. 106009, 2021.
[5] S. Sabour, G. Frosst, and N. Hinton, "Dynamic routing between capsules," *Advances in Neural Information Processing Systems*, vol. 30, pp. 3859–3869, 2017.
[6] P. Afshar, K. N. Plataniotis, and A. Mohammadi, "Capsule networks in healthcare: A survey," *Journal of Medical Systems*, vol. 44, no. 2, p. 98, 2020.
[7] Y. Li et al., "Diabetic retinopathy diagnosis using capsule networks," *IEEE Access*, vol. 10, pp. 25000–25010, 2022.
[8] J. Y. Choi et al., "Explainable computer-aided diagnosis of diabetic retinopathy using a deep neural network and its visual interpretation," *Diagnostics*, vol. 10, no. 9, p. 711, 2020.
[9] A. Singh et al., "Explainable deep learning models in medical image analysis," *Journal of Imaging*, vol. 7, no. 9, p. 73, 2021.
[10] M. Jampour et al., "Capsule networks for microaneurysm detection and explainability in diabetic retinopathy," *Medical Image Analysis*, vol. 75, p. 102260, 2022.
[11] 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.
[12] Kaggle, "APTOS 2019 Blindness Detection Dataset," 2019. [Online]. Available: https://www.kaggle.com/competitions/aptos2019-blindness-detection
[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.
[15] G. Litjens et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details