Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 23s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Exploring Explainable Artificial Intelligence Techniques for Diabetic Retinopathy Detection using Grad-CAM and VGG16 Framework

Bhaskar Marapelli[1], Ch Anil Carie[2], Dr. Ashish[3], Dr. Bechoo Lal[4], Dr. K Aruna Bhaskar[5]

[1]Department of Computer science and Engineering, Koneru Laxmaiah Education Foundation, Vaddeswaram, Guntur, India.
Email ID: bhaskarmarapelli@gmail.com

[2]Department of Computer science and Engineering, SRM university, Amaravathi, Inida.

Email ID: carieanil@gmail.com

[3]Department of Computer science and Engineering, Koneru Laxmaiah Education Foundation, Vaddeswaram, Guntur, India.
Email ID: dr.aashishsinha@gmail.com

[4]Research scholar, Department of computer science and engineering, Manipur International university- MIU Ghari, Airport Road, Imphal West Manipur,

Email ID: drblalpersonal@gmail.com

[5]Department of Computer science and Engineering, Koneru Laxmaiah Education Foundation, Vaddeswaram, Guntur, India.
Email ID: letter2arunbhaskar@gmail.com

Cite this paper as: Bhaskar Marapelli, Ch Anil Carie, Dr. Ashish, Dr. Bechoo Lal, Dr. K Aruna Bhaskar, (2025) Exploring Explainable Artificial Intelligence Techniques for Diabetic Retinopathy Detection using Grad-CAM and VGG16 Framework. *Journal of Neonatal Surgery*, 14 (23s), 344-354

ABSTRACT

Since diabetic retinopathy (DR) is one of the main causes of blindness in the world, avoiding vision loss requires early detection. Although deep learning models have demonstrated significant promise in the diagnosis of DR using retinal pictures, they frequently operate as "black boxes," which means that it is difficult to understand how they make decisions. Particularly when it comes to making important healthcare decisions, this lack of openness may erode confidence in AI-based medical solutions. Our research uses Explainable Artificial Intelligence (XAI) techniques to improve the interpretability of a deep learning model for DR detection in order to overcome this difficulty. We classify retinal images using a pre-trained convolutional neural network called VGG16.To enhance transparency, we incorporate Gradient-weighted Class Activation Mapping (GradCAM), a technique that generates visual heatmaps, highlighting the regions of the image that influenced the model's decision. By overlaying these heatmaps onto the original retinal images, we provide a clear, visual representation of what the model "sees" when making a classification. Our approach not only improves accuracy but also enhances trust in AI-based DR diagnosis. The accuracy of the VGG16 model was 72%, and Grad CAM successfully identified the crucial regions linked to DR. Deep learning models are simpler to incorporate into clinical practice thanks to these visual explanations that assist medical personnel in validating AI-generated predictions. Our research helps to improve patient outcomes by bridging the gap between interpretability and AI, which increases confidence among healthcare practitioners and makes AI-assisted medical diagnosis more transparent and trustworthy..

Keywords: *Deep learning, Diabetic retinopathy, Grad- CAM, VGG16 architecture, Visualize attention, XAI.*

1. INTRODUCTION

Approximately 30% of people with diabetes mellitus have diabetic retinopathy (DR), a microvascular complication resulting from the disease and a major cause of vision loss and blindness globally [1]. Early intervention to prevent vision loss depends on the automatic identification of DR. The identification of DR is aided by clinical, geometrical, and hemodynamic characteristics, emphasizing the importance of prompt screening [2]. Convolutional neural networks (CNNs), in particular, are deep learning (DL) models that have demonstrated remarkable performance in the automated detection of DR from fundus images [3]. However, the inherent 'black-box' nature of DL models often obscures the rationale behind their predictions, which hinders trust and acceptance by clinicians and patients alike [4]. Research highlights the importance of developing DL models that not only accurately detect DR but also provide clear and understandable insights into their decision-making process to address this issue [5]. Increasing the explainability of DL models can ultimately enhance the popularity and usefulness of automated DR detection systems in clinical practice, helping physicians and patients understand and trust the

diagnostic results [6]. The field of Explainable Artificial Intelligence (XAI) is expanding as a means of developing techniques that explain how AI models make decisions [7]. XAI approaches promote transparency, accountability, and trust by providing insights into how AI models arrive at specific judgments. XAI aids in identifying biases, validating model behavior, and understanding the importance of various features in model predictions. In medical imaging, XAI techniques offer the potential to increase trust in automated diagnoses and improve understanding of disease pathology. Grad-CAM is a prominent XAI visualization technique that generates heatmaps showing the image regions the model considers most significant for its classification decisions [8]. Grad-CAM improves model interpretability by highlighting important areas in the decision-making process, especially in medical imaging applications. Its ability to provide interpretable visualizations enhances the trustworthiness and transparency of deep learning models [9].

This study aims to tackle a crucial challenge in auto- mated Diabetic Retinopathy (DR) detection—namely, the need for transparency and interpretability in deep learning models. While convolutional neural networks (CNNs) have demonstrated remarkable accuracy in image classification tasks, their ”black-box” nature often leaves clinicians and users questioning how and why certain predictions are made. To address this, our research integrates the VGG16 architecture with Gradient-weighted Class Activation Mapping (Grad-CAM), providing visual explanations of the model’s decision-making process. VGG16 is a widely recognized and extensively used pre- trained CNN model originally developed for large-scale image recognition tasks. Its deep architecture, composed of 16 layers, has been fine-tuned and successfully applied across various medical imaging domains. For instance, VGG16 has been utilized to detect bone abnormalities [10], diagnose brain tumors [11], and classify Alzheimer’s disease using magnetic resonance imaging (MRI) scans [12]. Its proven effectiveness and generalization capabilities make it a strong candidate for DR detection tasks as well.

In our approach, we employ Grad-CAM to generate heatmaps that highlight the regions of retinal images the model focuses on when making predictions. These visual explanations help to demystify the internal workings of CNN by showing which parts of the image contributed most to the final decision. This layer of interpretability is crucial, especially in medical contexts, where understanding the rationale behind an automated diagnosis can help build trust among healthcare professionals and patients. By combining VGG16 with Grad-CAM, our work not only aims to maintain high diagnostic accuracy but also enhance the explainability and accountability of DR detection models. Ultimately, this approach supports more informed decision- making in clinical settings and paves the way for broader acceptance and integration of AI-powered tools in medical diagnostics.

The structure of this paper is as follows: Section 2 discusses Related Work, providing a review of relevant literature to establish context and identify research gaps. Section 3 explains the Methodology, detailing the approach, theoretical framework, data collection, and analytical methods used. In Section 4, the Experimental Setup is described, outlining the equipment, materials, software, and configurations employed. Section 5 presents the Results, which include quantitative and qualitative findings, supported by data analysis and visualization. Section 6 offers a thorough discussion, interpreting results, comparing them to existing literature, addressing limitations, and proposing future research directions. Finally, Section 7 summarizes the Conclusion, highlighting key findings, contributions, practical implications, and potential avenues for further research

2. RELATED WORK

According to [13], diabetic retinopathy (DR) is a dangerous complication of diabetes that affects the retina’s blood vessels. If not promptly identified and treated, it may result in permanent vision loss. For individuals with diabetes, early screening and monitoring are essential to reducing the risk of visual impairment. For DR detection and classification, traditional machine learning algorithms have been used alongside feature extraction methods; however, their accuracy remains subpar. Deep learning (DL) models, such as convolutional neural networks (CNNs), have revolutionized computer vision tasks, including DR detection. CNNs learn hierarchical features directly from raw image data, making them well-suited for complex visual tasks. In particular, transfer learning, which involves fine-tuning pre- trained models on specific tasks, achieves high accuracy with limited labeled data. It leverages previously learned knowledge and representations to adapt quickly to new tasks, proving invaluable when labeled data is scarce.

The research presented in [10] highlights the VGG16 model, developed by the Visual Geometry Group at the University of Oxford, which is renowned for its outstanding performance in image classification tasks. The work in [14] mentions that DL models, including VGG16, are often regarded as ’black boxes’ because their decision-making process is hidden within thousands of simulated neurons and interconnected layers. While these models achieve impressive accuracy, understanding their internal reasoning remains challenging due to the complexity of the many parameters and non-linear activation functions employed. In [13], researchers explored XAI methods to address the opaque nature of DL models. Specifically, they investigated Grad- CAM, an XAI technique that identifies the image regions most influential in predicting a particular class. In [15], the authors attempted to interpret black-box models to improve transparency and predictability, making them valuable for mission-critical applications in various domains.

Transfer learning with pre-trained models like VGG16, as discussed in [16], [17], [18], enables the use of knowledge learned from large datasets. By fine-tuning VGG16 on custom datasets (such as DR images), accurate DR classification can be

achieved. This approach proves especially beneficial in medical image analysis, where annotated datasets are limited, significantly improving classification performance. XAI techniques, such as Grad-CAM, provide visual explanations of DL model predictions, allowing clinicians and researchers to understand the decision-making process and identify regions of interest in medical images [19]. The authors in [19], [20] explain that Grad-CAM promotes transparency, builds trust, improves model robustness, and ensures patient safety in medical applications. Grad-CAM serves as a baseline method for highlighting critical image regions involved in DL model decision-making, enhancing interpretability and trust in medical applications such as DR diagnosis. Grad-CAM has been widely used to enhance the interpretability of deep learning models across various architectures, including CNNs, Vision Transformers, and Swin Transformers. By generating heatmaps that highlight important regions in input images, the technique helps visualize how models make decisions [21].

In medical imaging, Grad-CAM has played a crucial role in improving the understanding of model predictions, particularly in brain tumor detection and classification. It provides valuable insights that go beyond conventional evaluation metrics, making AI-driven diagnoses more transparent and reliable [22]. This technique has been effectively applied in COVID-19 diagnosis using chest X-rays. By offering clear visual explanations of model predictions, Grad- CAM has contributed to improving both the accuracy and interpretability of AI-assisted medical assessments. Models utilizing Grad-CAM have reported accuracy rates as high as 99.09%, outperforming existing models and demonstrating potential for reliable clinical deployment [23].

3. METHODOLOGY

The effectiveness of utilizing deep features extracted from the VGG16 model has been demonstrated in tasks like brain tumor detection [24], showcasing the robustness and versatility of the VGG16 architecture in various image analysis tasks. By using the VGG16 model—well-known for its exceptional performance in image classification—we can enhance the interpretability of the model's predictions on diabetic retinopathy images. This approach aligns with the current trend in literature, where deep learning models, such as CNNs, are employed for the early detection and classification of diabetic retinopathy ([25]; [26]). To apply Grad-CAM to diabetic retinopathy images using the VGG16 model, we utilized the pre-trained VGG16 model's convolutional layers to visualize the regions crucial for the model's predictions. Grad-CAM generates a heatmap that highlights the important areas of the input image that influence the model's decision-making [27].

- Transfer Learning

Transfer learning (TL) has emerged as a powerful technique in machine learning, enabling the reuse of models developed for one task to improve performance on a related but different task. This approach has revolutionized various do- mains, from medical imaging to natural language processing, by addressing challenges such as data scarcity, computational inefficiency, and domain adaptation. This response provides a comprehensive exploration of the applications, strategies, and implications of transfer learning, drawing insights from recent research papers. Transfer learning is a pivotal machine learning technique that enhances model performance by leveraging knowledge from one task to improve another related task. Transfer learning involves using a pre-trained model (source domain) to kickstart a new model for a different but related task [28]. It can significantly reduce the amount of data and time required to train models, as seen in applications like image classification and control systems [29],[28]. Transfer learning can effectively transfer control logic, ensuring behavioral equivalence between systems without needing extensive verification [28]. Transfer learning in developing a model for Alzheimer's detection from MRI images, achieved over 88% accuracy with minimal data [29].

A common transfer learning strategy involves using pretrained models as feature extractors by freezing the initial layers, which capture general low-level features, and fine- tuning the later layers to adapt to the specific target task. This approach leverages the generalized knowledge learned from large datasets like ImageNet, enabling effective learning even with limited data. It is widely used in image recognition tasks due to its efficiency, faster convergence, and improved performance [36]. To implement transfer learning a deep learning technique a model is trained on a large dataset and is reused for a different but related task with a smaller dataset. It follows a three-step process: pre-training, feature extraction, and fine-tuning. First, a pre-trained model such as ResNet (for images) or BERT (for NLP) is trained on a massive dataset like ImageNet or a large text corpus. These models learn general features such as edges, shapes, or language patterns. In the feature extraction phase, the knowledge gained by the pre trained model is transferred to the new task. Instead of training a model from scratch, which requires extensive data and computing power, the already learned weights and feature representations are utilized. This approach significantly reduces training time and improves performance, especially when the target dataset is small. In the fine-tuning phase, the transferred model is slightly adjusted by training it on the new dataset. This step ensures that the model adapts to the specific task while retaining useful knowledge from the original training. Transfer learning is widely used in applications like medical imaging, speech recognition, and NLP, where labeled data is limited but leveraging pre-trained models improves accuracy and efficiency.

- VGG-16 Model

The VGG16 model is a deep convolutional neural network architecture that has been widely utilized in many computer vision tasks. It was first trained on the ImageNet dataset with 1000 classifications [37]. The Visual Geometry Group at the

prestigious University of Oxford presented the CNN model that is represented by the architectural architecture of VGG-16. This model consists of 16 layers, including 3 fully connected layers, convolutional layers, and max-pooling layers. Renowned for its straightforwardness and efficiency in the realm of large-scale image classification, VGG-16 has been deployed across a spectrum of domains, ranging from the identification of weeds in corn fields to the categorization of fruits and vegetables, recognition of Chinese herbal medicines, and detection of brain tumors. Various research endeavors have showcased the superior accuracy and performance of VGG-16, surpassing alternative models in select scenarios. Notably, in the domain of brain tumor detection, VGG-16 achieved a commendable accuracy rate of 94% after the optimization of hyperparameters, thereby exhibiting robust sensitivity and specificity. Collectively, VGG-16 emerges as a potent instrument within the domain of deep learning, specifically tailored for a diverse array of image recognition tasks.

VGG-16 is composed of a series of convolutional layers with max-pooling layers alternating with Rectified Linear Unit (ReLU) activations, which gradually decrease the spatial dimensions. Subsequently, the architecture transitions to a series of fully connected layers culminating in the ultimate output layer.

Convolutional Layer The output of a convolutional layer is obtained by applying the convolution operation followed by an activation function:

$$X^{(l)} = f^{(l)}(Conv(X^{(l-1)}, W^{(l)}) + b^{(l)}) \quad (1)$$

Max Pooling Layer Max pooling reduces the spatial dimensions of the input feature maps:

$$X^{(l)} = MaxPool(X^{(l-1)}) \quad (2)$$

Fully Connected Layer The fully connected layer's out- put is produced by applying an activation function, adding bias, and multiplying the input by weights:

$$X(l) = f(l)(X(l-1)\, W(l) + b(l)) \quad (3)$$

In the above equations (1),(2),(3) $X^{(l-1)}$ is the the input to layer l, $X^{(l)}$ is the output of layer l, $W^{(l)}$ is the weights of layer l, $b^{(l)}$ is the bias of layer l, $f^{(l)}$ is the activation function of layer l, $Conv()$ is the convolution operation, $MaxPool()$ is the max pooling operation.

- Grad-CAM

Grad-CAM is an interpretability technique used to explain the predictions made by deep neural networks, particularly CNNs. Grad-CAM aims to identify the important regions in an input image that significantly influences the classification score produced by the network. It provides insights into why a particular class was predicted by highlighting the relevant parts of the input [38], [39]. Grad-CAM, a technique used in image analysis, helps understand which image regions are most crucial for a model's classification. It achieves this by analyzing the gradients of the classification score (the model's confidence in a particular class) concerning the final layer's feature map. Intuitively, larger gradients indicate that modifying these features significantly impacts the classification score. Therefore, the areas with high gradients in the resulting heatmap correspond to the image regions most influential for the model's decision.

Fig. 1. Transfer Learning

TABLE I Comparative Analysis of Transfer Learning Applications

Domain	Key Applications	Challenges and Strategies
Medical Imaging	Disease diagnosis, image segmentation, mutation detection	Mismatch between natural and medical images; novel pre-training approaches on unlabeled medical data [30][31]
NLP	Text classification, question	Sensitivity to dataset features and

	answering, language translation	domain-specific fine-tuning requirements [32]
Robotics	Control policy adaptation, simulation-to-real-world transfer	Ensuring behavioral guarantees and addressing domain shifts [33][34]
Healthcare (EHRs)	Detection of treatment disparities, hospital stay prediction	Unbalanced data and domain adaptation using optimal transport [35]

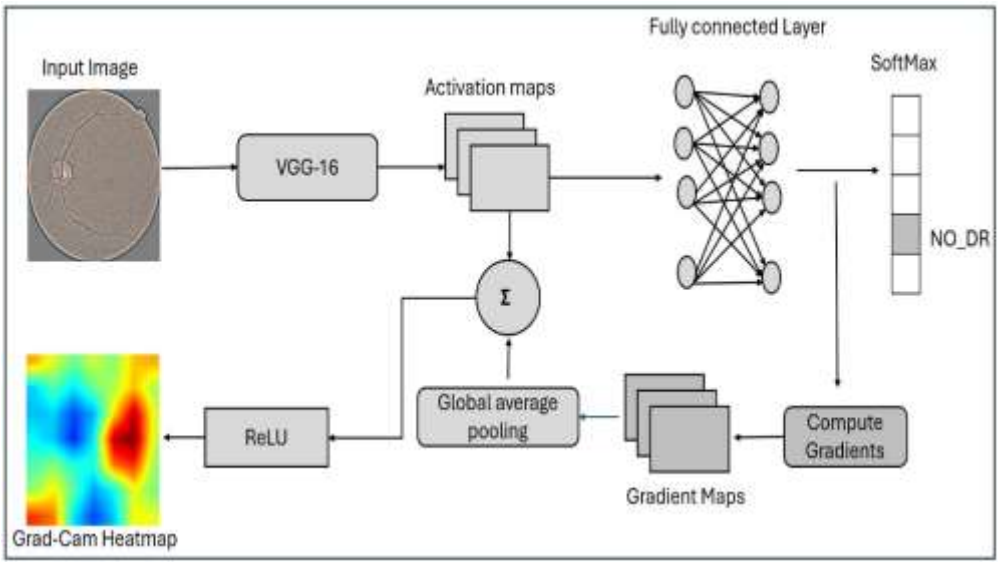

Fig. 2.Grad-Cam Heat map Generation

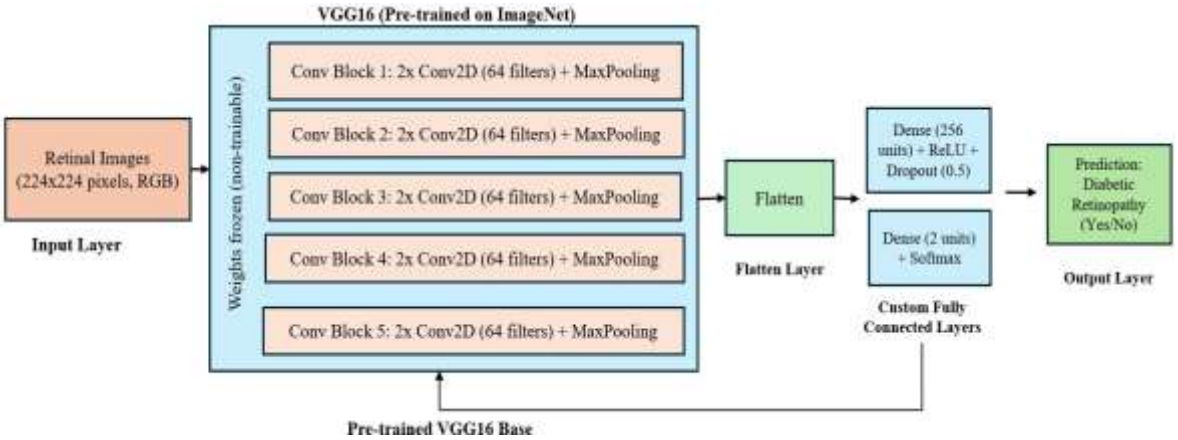

Fig. 3. Transfer Learning with VGG16 for Diabetic Retinopathy

Activation Map:

Let A_{ijk}^{l} denote the activation of the k-th feature map at position (i,j) in the last convolutional layer l.

Gradient Score:

The gradient of the class score C with respect to the activation map is calculated as:

$\alpha_k^l = (1 / Z) \times \sum_{(ij)} (\partial C / \partial A_{ijk}^l)$

where Z is the normalization factor (typically $Z = i \times j$).

Weighted Sum:

The Grad-CAM map is obtained by taking a weighted sum of the activation maps, using the gradient scores:

$\text{Grad-CAM} = \text{ReLU}(\sum_k \alpha_k^l \times A_{ijk}^l)$

where ReLU is the rectified linear unit function.

Final Visualization:

Grad-CAM highlights the regions where the network's decision is most influenced by the features, providing insights into the CNN's decision-making process.

- Experimental Setup

Dataset

For this Experimental setup, a dataset of retina scan images is used to detect diabetic retinopathy which is freely available in the Kaggle repository. The images are preprocessed with Gaussian filtering and resized to 224x224 pixels for compatibility with pre-trained deep-learning models. The dataset is categorized into five severity/stage classes: No_DR, Mild, Moderate, Severe, and Proliferate_DR, each represented by a respective directory containing images. These classifications are based on the severity of diabetic retinopathy.

Model Building and Experiment

We used Grad-CAM in our experimental setup to identify the critical areas in the image that significantly affect the predictions of a trained VGG16 model. To extract features from input images, we first imported a pre-trained VGG16 model from the Keras library and used it as a feature extractor. Then, an Artificial Neural Network (ANN) model with a fully connected layer connected to an output layer with five output neurons was built using these retrieved features. We trained this combined VGG16-ANN model using diabetic retinopathy data. Throughout the training, we kept the VGG16 weights unchanged. After training we loaded an image and preprocessed it to conform to the input size requirements of the model. Subsequently, predictions were made on this preprocessed image using the trained model, and the predicted class label was determined. Next, we extracted the activation map of the last convolutional layer to identify which regions of the image were activated by the model. To highlight the regions essential for the model's prediction, we created Grad-CAM by computing the gradients of the anticipated class concerning this activation map. This heatmap was then superimposed onto the original image to visualize the regions contributing the most to the model's decision-making process. For visualization purposes, we plotted the original image, the Grad-CAM heatmap, and the overlay image. This allowed us to understand the reasoning behind the model's prediction; sample output is shown in Figure 3. This all-encompassing strategy made it easier to comprehend how the model made decisions and gave its predictions important interpretability.

4. RESULTS AND DISCUSSION

The prepared VGG16_ANN models trained for 10 epochs, each consisting of batches. Throughout the training, accuracy improves from 53.70% to 73.34%, while loss decreases from 2.9404 to 0.7363. Validation accuracy reaches 72.01% at the end of training, with a corresponding loss of 0.7698. These metrics indicate progressive improvement in the model's performance on both training and validation data. Figures 8 and 9 show the Loss and Accuracy Graphs for 10 epochs.

After training the model to generate Grad-CAM visualizations, the activation function of the last layer is changed to linear and the heatmap is generated with the last layer of VGG16. From the table, it is evident that the model’s performance improved significantly during the training process. The training loss, which began at 1.8, dropped to 0.7 by the 9th epoch. This decrease indicates that the model has become better at reducing errors in the training data. Likewise, the validation loss also fell, going from 1.0 to 0.7, which points to the model’s improved ability to generalize new data. The training accuracy increased from 62% to 74%, demonstrating the model’s enhanced accuracy in predicting the training data. Similarly, validation accuracy improved from 64% to 74%, showing that the model's predictions on unseen data became more reliable as training progressed. These changes suggest that the model was able to learn effectively and generalize well, resulting in better performance on both training and validation datasets. To visualize a comparison between the original images and the areas of interest identified by the model, the technique known as Grad-CAM is utilized. These regions—often referred to as "attention regions"—are emphasized to shed light on the aspects of the picture that most strongly influence the model's predictions.

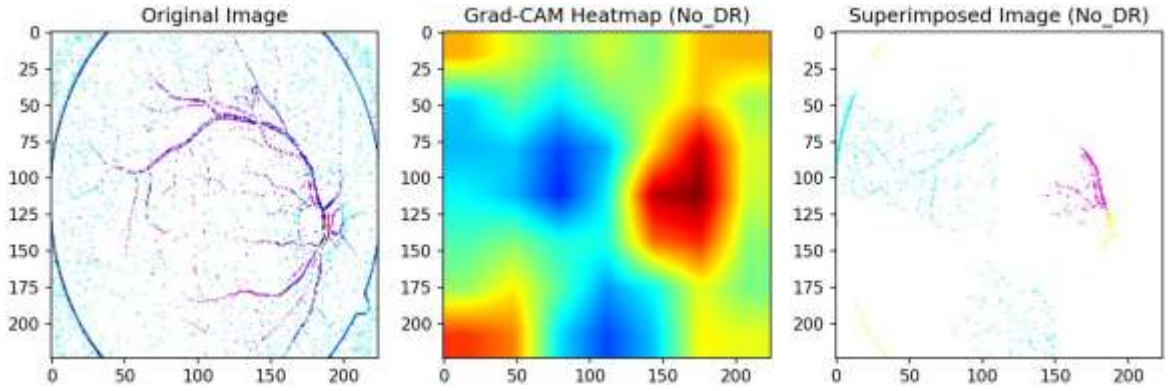

Fig. 4. Sample Prediction By Grad-Cam

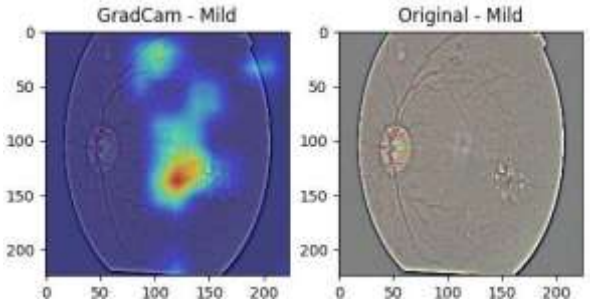

Fig. 5. GradCam Prediction on Mild Diabetic Retinopathy Image

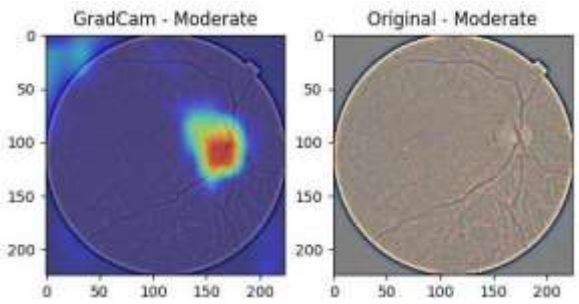

Fig. 6. GradCam Prediction on Moderate Diabetic Retinopathy Image

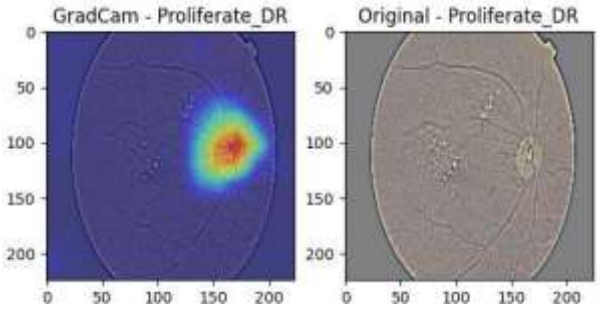

Fig. 7. GradCam Prediction on Proliferate Diabetic Retinopathy Image

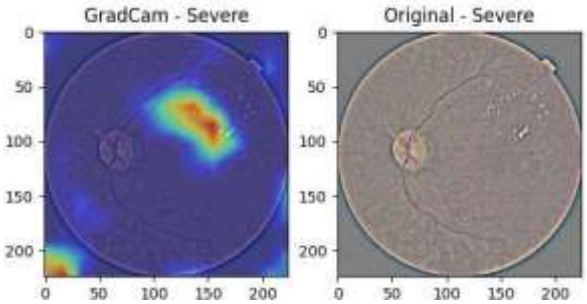

Fig. 8. GradCam Prediction on Severe Diabetic Retinopathy Image

In this process, Grad-CAM produces heatmaps that show the relative importance of various image regions for the categorization decision made by the model. These heatmaps are then overlaid onto the original images using the Matplotlib library, a versatile plotting tool in Python commonly used for data visualization. By overlaying the Grad-CAM heatmaps onto the original images, we generate visuals that highlight the regions of the image that the model utilizes as the basis for diabetic retinopathy predictions.

The figures referenced as Figures 4, 5, 6, and 7 depict these visualizations. Each figure displays an original image alongside its corresponding Grad-CAM heatmap overlay, highlighting the regions deemed crucial by the model for predicting diabetic retinopathy. These visualizations improve the interpretability and comprehension of the deep learning network's predictions and provide insightful information about how the algorithm makes decisions.The training process spans 10 epochs, during which both accuracy and validation accuracy steadily improve, while loss steadily decreases. This indicates the model's progressive enhancement in learning and generalization. Following the training phase, Grad-CAM visualizations are generated to interpret the model's decisions. By overlaying heatmaps onto original images, the regions crucial for predictions are highlighted.

From the table, it is evident that the model's performance improved significantly during the training process. The training loss, which began at 1.8, dropped to 0.7 by the 9th epoch. This decrease indicates that the model has become better at reducing errors in the training data. Likewise, the validation loss also fell, going from 1.0 to 0.7, which points to the model's improved ability to generalize new data. The training accuracy increased from 62% to 74%, demonstrating the model's enhanced accuracy in predicting the training data. Similarly, validation accuracy improved from 64% to 74%, showing that the model's predictions on unseen data became more reliable as training progressed. These changes suggest that the model was able to learn effectively and generalize well, resulting in better performance on both training and validation datasets.

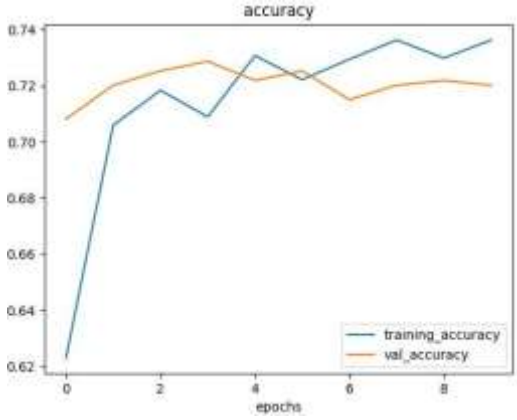

Fig. 9. Train and Test Loss graph for Vgg16_ANN model

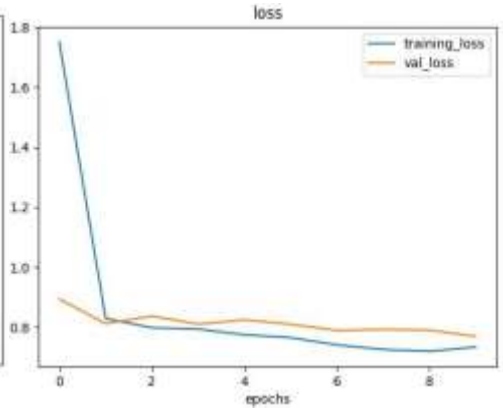

Fig. 10. Train and Test Accuracy graph for Vgg16_ANN model

TABLE II Accuracy, Loss comparison of the model

Epoch	Training Loss	Validation Loss	Training Accuracy	Validation Accuracy
1	1.8	1	0.62	0.64
2	1.5	0.9	0.65	0.66
3	1.2	0.85	0.68	0.68
4	1	0.8	0.7	0.7
5	0.9	0.78	0.71	0.71
6	0.85	0.75	0.72	0.72
7	0.8	0.73	0.73	0.73
8	0.75	0.72	0.73	0.73
9	0.72	0.71	0.74	0.74
10	0.7	0.7	0.74	0.74

The goal of this work is in alignment with the increasing demand for interpretability and transparency in artificial intelligence (AI) systems, especially when it comes to deep learning and image classification applications. DL models, such as VGG16, often function as powerful black-box systems, capable of making complex decisions but lacking transparency in how those decisions are made. This opacity might make it difficult to diagnose possible biases or errors and to comprehend the internal

workings of the model. Researchers and practitioners can obtain significant insights into the model's decision-making process by employing techniques such as Grad-CAM to visualize the regions of input images that contribute most to the model's predictions.

These visualizations provide a window into the "thought process" of the model, highlighting the areas of the image that, in this scenario, have the most influence on the classification decisions made with diabetic retinopathy. Understanding the model's behavior in this way serves multiple purposes. Firstly, it facilitates interpretability, allowing stakeholders to comprehend why the model made a particular prediction. This insight is crucial in domains where decisions impact human lives, such as healthcare, as it enables clinicians to trust and verify the model's decisions. Also, it can help guide modifications and enhancements by pointing up possible biases or errors in the architecture of the model or training set.

5. CONCLUSION

This work demonstrates Transfer learning with pre-trained model VGG16 for image classification tasks and XAI techniques, such as Grad-CAM, gives predictions from the DL model visual explanations. This approach can help com- prehend and improve the model's performance by offering insightful information about the model's decision-making process. DL models have shown promise in DR detection, yet their" black-box" nature often hampers interpretability and trust in their decision-making processes. To address this challenge, we explored the application of XAI techniques. This work provides a comprehensive approach to training CNNs, interpreting their predictions, and enhancing their understanding of deep learning models' behavior, functioning as a useful tool for machine learning applications and jobs involving image classification. This research contributes to the ongoing efforts in leveraging advanced machine learning methods for medical image analysis while addressing the critical need for interpretability and transparency in health- care AI systems.

Acknowledgment

Not Applicable

REFERENCES

[1] J. Zhou and B. Chen, "Retinal cell damage in diabetic retinopathy," Cells, vol. 12, no. 9, p. 1342, 2023.

[2] B. P. Makala, M. Kumar, and B. Sirisha, "Survey on automatic detection of diabetic retinopathy screening," in 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART). IEEE, 2022, pp. 1214–1220.

[3] R. S. Gound, B. M. Sundaram, S. K. BV, P. A. Azmat, M. N. U. Habib, and A. Garg, "Drs-unet: A deep learning approach for dia- betic retinopathy detection and segmentation from fundus images," in 2023 4th International Conference for Emerging Technology (INCET). IEEE, 2023, pp. 1–5.

[4] A. Bianchi, A. Di Marco, F. Marzi, G. Stilo, C. Pellegrini, S. Masi, Mengozzi, A. Virdis, M. S. Nobile, and M. Simeoni, "Trustworthy machine learning predictions to support clinical research and deci- sions," in 2023 IEEE 36th International Symposium on Computer- Based Medical Systems (CBMS). IEEE, 2023, pp. 231–234.

[5] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," Ophthalmology, vol. 124, no. 7, pp. 962–969, 2017.

[6] N. K. Baskaran and T. Mahesh, "Performance analysis of deep learning based segmentation of retinal lesions in fundus images," in 2023 Second International Conference on Electronics and Renewable Systems (ICEARS). IEEE, 2023, pp. 1306–1313.

[7] M. Garouani and M. Bouneffa, "Unlocking the black box: Towards interactive explainable automated machine learning," in International Conference on Intelligent Data Engineering and Automated Learning. Springer, 2023, pp. 458–469.

[8] Z. Chen and Q. Sun, "Extracting class activation maps from non- discriminative features as well," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3135–3144.

[9] Y. Liao, Y. Gao, and W. Zhang, "Feature activation map: Visual explanation of deep learning models for image classification," arXiv preprint arXiv:2307.05017, 2023.

[10] A. M. Barhoom, M. R. J. Al-Hiealy, and S. S. Abu-Naser, "Bone abnormalities detection and classification using deep learning-vgg16 algorithm," Journal of Theoretical and Applied Information Technol- ogy, vol. 100, no. 20, pp. 6173–6184, 2022.

[11] P. Gayathri, A. Dhavileswarapu, S. Ibrahim, R. Paul, and R. Gupta, "Exploring the potential of vgg-16 architecture for accurate brain tu- mor detection using deep learning," Journal of Computers, Mechanical and Management, vol. 2, no. 2, pp. 23 056–23 056, 2023.

[12] F. A. Torghabeh, Y. Modaresnia, and M. M. Khalilzadeh, "Effective- ness of learning rate in dementia severity

prediction using vgg16," Biomedical Engineering: Applications, Basis and Communications, vol. 35, no. 03, p. 2350006, 2023.

[13] S. Goel, S. Gupta, A. Panwar, S. Kumar, M. Verma, S. Bourouis, and M. A. Ullah, "Deep learning approach for stages of severity clas- sification in diabetic retinopathy using color fundus retinal images," Mathematical Problems in Engineering, vol. 2021, pp. 1–8, 2021.

[14] T. Qamar and N. Z. Bawany, "Understanding the black-box: towards interpretable and reliable deep learning models," PeerJ Computer Science, vol. 9, p. e1629, 2023.

[15] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain, "Interpreting black-box models: a review on explainable artificial intelligence," Cognitive Computation, vol. 16, no. 1, pp. 45–74, 2024.

[16] M. K. Singh and B. Kumar, "Fine tuning the pre-trained convolutional neural network models for hyperspectral image classification using transfer learning," in Computer Vision and Robotics: Proceedings of CVR 2022. Springer, 2023, pp. 271–283.

[17] A. Abhishek, S. D. Deb, R. K. Jha, R. Sinha, and K. Jha, "Classi- fication of leukemia using fine tuned vgg16," in 2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT). IEEE, 2023, pp. 1–5.

[18] P. Borugadda, R. Lakshmi, and S. Sahoo, "Transfer learning vgg16 model for classification of tomato plant leaf diseases: A novel approach for multi-level dimensional reduction." Pertanika Journal of Science & Technology, vol. 31, no. 2, 2023.

[19] J. Stodt, M. Madan, C. Reich, L. Filipovic, and T. Ilijas, "A study on the reliability of visual xai methods for x-ray images," Healthcare Transformation with Informatics and Artificial Intelligence, vol. 305, p. 32, 2023.

[20] S. K. Mandala, "Xai renaissance: Redefining interpretability in medical diagnostic models," arXiv preprint arXiv:2306.01668, 2023.

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and Batra, "Grad-cam: visual explanations from deep networks via gradient-based localization," International journal of computer vision, vol. 128, pp. 336–359, 2020.

[22] K. Kumar and K. Jyoti, "Recent advancements in grad-cam and variants: Enhancing brain tumor detection, segmentation, and classifi- cation," 2024.

[23] A. Bhatnagar, "Attention-driven convolutional neural network for accurate covid-19 diagnosis in medical imaging with gradcam-based interpretability," in 2024 International Conference on Decision Aid Sciences and Applications (DASA). IEEE, 2024, pp. 1–7.

[24] J. D. Bodapati, A. Vijay, and N. Veeranjaneyulu, "Brain tumor detec- tion using deep features in the latent space." Inge´nierie des Syste`mes d'Information, vol. 25, no. 2, 2020.

[25] Q. H. Nguyen, R. Muthuraman, L. Singh, G. Sen, A. C. Tran, B. P. Nguyen, and M. Chua, "Diabetic retinopathy detection using deep learning," in Proceedings of the 4th international conference on machine learning and soft computing, 2020, pp. 103–107.

[26] D. Pruthviraja, B. Anil, and C. Sowmyarani, "Efficient local cloud- based solution for diabetic retinopathy detection," International Jour- nal of Web-Based Learning and Teaching Technologies (IJWLTT), vol. 16, no. 3, pp. 39–46, 2021.

[27] Q. Zhang, L. Rao, and Y. Yang, "A novel visual interpretability for deep neural networks by optimizing activation maps with perturba- tion," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 4, 2021, pp. 3377–3384.

[28] M. M. Islam, "The impact of transfer learning on ai performance across domains," Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023, vol. 1, no. 1, 2024.

[29] M. Zoric´, M. S˘tula, I. Markic´, and M. Braovic´, "Transfer learning in building neural network model case study," in 2024 International Conference on Software, Telecommunications and Computer Networks (SoftCOM). IEEE, 2024, pp. 1–6.

[30] L. Alzubaidi, M. Al-Amidie, A. Al-Asadi, A. J. Humaidi, O. Al- Shamma, M. A. Fadhel, J. Zhang, J. Santamar´ıa, and Y. Duan, "Novel transfer learning approach for medical imaging with limited labeled data," Cancers, vol. 13, no. 7, p. 1590, 2021.

[31] M. Angriawan, "Transfer learning strategies for fine-tuning pretrained convolutional neural networks in medical imaging," Research Journal of Computer Systems and Engineering, vol. 4, no. 2, pp. 73–88, 2023.

[32] P. Banerjee and S. Kashyap, "Unlocking transfer learning's potential in natural language processing: An extensive investigation and evalu- ation," in 2024 International Conference on Advances in Computing Research on Science Engineering and Technology (ACROSET). IEEE, 2024, pp. 1–7.

[33] N. Jaquier et al., "Transfer learning in robotics: an upcoming break- through," A review of promises and challenges, 2023.

[34] A. Nadali, B. Zhong, A. Trivedi, and M. Zamani, "Transfer learning for control systems via neural simulation relations," arXiv preprint arXiv:2412.01783, 2024.

[35] W. Li, Y. P. Park, and K. D. Duc, "Transport-based transfer learning on electronic health records: Application to detection of treatment disparities," medRxiv, pp. 2024–03, 2024.

[36] A. Alshardan, N. Alruwais, H. Alqahtani, A. Alshuhail, W. S. Al- mukadi, and A. Sayed, "Leveraging transfer learning-driven convolu- tional neural network-based semantic segmentation model for medical image analysis using mri images," Scientific Reports, vol. 14, no. 1, p. 30549, 2024.

[37] Y. Zhao, H. Zhou, S. Lu, Y. Liu, X. An, and Q. Liu, "Human activity recognition based on non-contact radar data and improved pca method," Applied Sciences, vol. 12, no. 14, p. 7124, 2022.

[38] H. Kafri, M. Olivieri, F. Antonacci, M. Moradi, A. Sarti, and S. Gan- not, "Grad-cam-inspired interpretation of nearfield acoustic holography using physics-informed explainable neural network," in ICASSP 2023- 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–5.

[39] Y. Yang, R. Guo, S. Wu, Y. Wang, J. Zhang, X. Gong, and B. Zhang, "Decom–cam: Tell me what you see, in details! feature-level in- terpretation via decomposition class activation map," arXiv preprint arXiv:2306.04644, 2023.

.